Rails and innovation: Evidence from China*

Georgios Tsiachtsiras† Deyun Yin ‡ Ernest Miguelez §
Rosina Moreno ¶

September 13, 2021

Abstract

In this paper, we explore the effect of the High Speed Rail (HSR) network expansion in China on patent activity of cities, for the period 2008-2016. Using exogenous variation arising from a novel instrument, based on courier's stations during Ming dynasty, we find solid evidence that the opening of a HSR station is associated with an increase ina city's innovation activity. We report evidence that this relationship exists not only in the large urban centers but also among third-tier cities, more peripheral in the urban system of Chinese cities. Among the several reasons why HSR inter-city connectivity increases city innovation, we explore the role of inter-city technology diffusion. In par- ticular, we compute least-cost paths based on the speed of each HSR line and we docu- ment for the first time that an increase in a city's connectivity to other cities specializingin a specific technological field through the HSR network increases the probability for the city to specialize in the same technological field, which we interpret as evidenceof knowledge diffusion. We also find that the expansion of HSR network enhances cross-city co-patenting and the geographical mobility of inventors.

Keywords: high speed rail, innovation, diffusion, patent, specialization

JEL classification: 018, 030, 033

*We would like to express our gratitude to Ron Boschma, Riccardo Crescenzi, Giovanni Dosi, Francesco Lissoni, Alex Coad, Antonio Andreoni, Carolina Castaldi, Bart Los, Jouke van Dijk, Ruben Gaetani, Martijn Smit, Yuan Tian, Nicola Cortinovis, Michele Pezzoni, Zhiling Wang, Matteo Tubiana, Mathieu Steijn, Fernando Stipanicic, Julia Mazzei and Dongmiao Zhang for their valuable comments and suggestions. Furthermore, we would like to thank the participants of the EPIP 2021 PHD student workshop, the 18th International Schumpeter Society (ISS) Conference (online), the 9th summer school on "Knowledge Dynamics, Industrial Evolution, Economic Development" (KID 2021, University of Cote d'Azur), the 34th ERSA Summer School (University of Groningen, online), the Economic Geography Seminar (University of Utrecht, Online), the 10th European Meeting of the Urban Economics Association (Copenhagen Business School, Online), the 8th International PhD Workshop in Economics of Innovation, Complexity and Knowledge (University of Turin, Online) and the 5th Summer school on Data and Algorithms for ST&I studies (KU Leuven, Online) for their comments. Any remaining errors are ours.

†Tsiachtsiras: Ph.D. candidate, University of Barcelona School of Economics; Department of Statistics, Econometrics, and Applied Economics, AQR-IREA, Barcelona 08034, Spain. email: gtsiachtsiras@ub.edu

‡Yin: School of Economics and Management, Harbin Institute of Technology, Shenzhen, Guangdong Province, China. email: yindeyun@hit.edu.cn

§Miguelez: GREThA-UMR CNRS 5113, Université de Bordeaux, Avenue Léon Duguit, 33608 Pessac, France and AQR-IREA, University of Barcelona, Spain. email: ernest.miguelez@u-bordeaux.fr

¶Moreno: AQR-IREA, University of Barcelona, Barcelona, Spain. email: rmoreno@ub.edu

1 Introduction

Over the last decades, China has become a science and technology powerhouse, and ambitions to become an innovation global leader by mid of the 21st century (Veugelers, 2017). Since 2011, the China National Intellectual Property Administration (CNIPA) has outnumbered other patent offices, like the JPO or the USPTO, in terms of the applications it receives (WIPO, 2012). If one concentrates on the international innovations actually produced in China's territory (patent families produced by Chinese firms and inventors, and with at least one extension abroad), China has recently overcome the Republic of Korea, reaching 14% of all patent production, not far from other big patent producers (US, 21.1%; Japan, 21%; Europe, 23.9%) (WIPO, 2019). Yet, as in many other countries, the internal spa-tial distribution of innovation is highly skewed. While Beijing, Shanghai, and Shenzhen- Hong Kong concentrated 36.5% of all internationally-oriented patents produced in Chinain 1991-1995, these three leading innovation hotspots accounted for 52.2% of all interna- tional patents in the period 2011-2015 (WIPO, 2019).

Innovation is above all an urban phenomenon (Carlino and Kerr, 2014; Florida et al., 2017). All sorts of agglomeration economies make individuals, especially high-skilled ones, more productive in cities (Lucas, 1993). Agglomerations also improve the employer-employee match quality, in turn affecting labor market efficiency and productivity (Moretti, 2019). Larger market access in big cities increases innovator incentives to develop their ideas (Ellison et al., 2010). Cities allow interacting and learning from peers, favouring localized knowledge spillovers (Jaffe et al., 1993). The generation of new ideas relies upon the integration and (re)combination of various knowledge bases, which are usually distributed unevenly across different regions (Fleming, 2001; Strambach and Klement, 2012; Teixeira et al., 2008). Complex activities are concentrated in big cities (Balland et al., 2020) while rural areas are lagging behind (Rodríguez-Pose, 2018).

Transport infrastructure investments in cities, and across them, affect how people and firms interact, and may affect the geography of innovation and other economic outcomes (Agrawal et al., 2017). They allow longer commuting, favoring better quality employer-employee matches and, in general, labor market pooling. They drastically changes market size and market opportunities of local innovators. Scientists and inventors see increased their opportunities to collaborate and build new and larger teams, generating more ideas (Wuchty et al., 2007; Dong et al., 2020). Finally, they allow connecting locations far apart (Cao et al., 2021) and the recombination of knowledge pieces across the space (Owen-

Smith and Powell, 2004).

In this paper, we focus on the effect of HSR intercity connectivity on knowledge creation and diffusion for the period 2008-2016. We find solid evidence that the opening of a HSR station is associated with an increase in the number of patents per capita of a city. We further argue that this relationship exists not only in the large urban centers but also among second and third tier cities, more peripheral in the urban system of Chinese cities. We then investigate the extent to which HSR rollout affects knowledge diffusion, too. First, we test the effect of the expansion of HSR network on cross-city patent co-applications and inventors' mobility. These channels allow a broader reach of otherwise localized knowledge spillovers (Jaffe et al., 1993). Second, we test whether the reduction in transportation cost, due to infrastructure improvement, affect knowledge diffusion channels, as proxied by knowledge diversification. In particular, we borrow from the branching literature the idea of technological diversification (Hidalgo et al., 2007; Hausmann et al., 2007; Essletzbichler, 2015; Rigby, 2015; Boschma, 2017), and study the probability of a city to specialize in a new technological field, as a function of the specialization patterns of the cities to which the city connects through HSR, other things being equal. According to the mentioned literature, the probability that a city diversifies in a new technology (develops comparative advantage) increases with the presence of related technologies in that city, in line with the principle of relatedness, (Guo and He, 2017; He et al., 2017; Hidalgo et al., 2007, 2018; Gao et al., 2021; Zhu et al., 2019). Recent examples of this literature include, among others, Balland et al. (2019), who report evidence that relatedness and knowledge complexity affects the probability of a region to specialize in a given technological field. Petralia et al. (2017) investigate how the proximity of countries' existing capabilities to every technology is associated with the probability to specialize in a new technological field. Finally, Rigby (2015) documents that technological proximity has an impact on the specialization of US cities. None of these studies, however, address the issue of diversification thanks to the connections to the outside world.

This paper also contributes to the literature linking transport infrastructure, such as the HSR, and economic outcomes in China, which is now vast (Duranton et al., 2013). During the last decade, there has been a growing interest on linking the effect of HSR in China to all sorts of economic outcomes. Zheng and Kahn (2013) argue that HSR fosters real estate prices in nearby secondary cities by offering households and firms a larger menuof location alternatives. Transportation improvements facilitate individuals accessing the big urban centers without living within their boundaries. Qin (2017) explores the effect of

HSR on GDP in China and find that the economic activities move from peripheral counties to the urban core when the transport cost of people connecting between urban areas is reduced. Gao et al. (2018) confirm the findings of Qin (2017) using a county-level panel dataset of China's Yangtze River region. Lin (2017) finds that access to HSR increases GDP and urban employment. After a city gets connected to the HSR network, it witnesses an 18% increase in the number of passengers travelling by train and a 9.6% increase in the number of passengers travelling by any forms of transportation. Finally, Ke et al. (2017) investigate the effect of HSR projects on economic growth of targeted city nodes. The authors document that the responses to HSR network are heterogeneous and depend from the location, route, and region of the cities. For instance, cities along the Hu-Ning Segment, the Yong-Tai-Wen-Fu-Xia Segment, and within the Hunan province along the Wu-Guang HSR experience positive shocks on their economic growth. Building on the previous paper, Dong (2018) indicates that employment growth also varies with regard to the different industries. For instance, HSR affects retail/wholesale and hotel/food industries, while leaves unaffected the other sectors. Hong and Su (2019) show that improvements in the accessibility of Chinese cities have a crucial role in the formation of economic linkages among these cities. Zou et al. (2019) argue that market access, based on HSR network, contributes to the economic growth of Chinese cities. On the other hand, Jin et al. (2020) present evidence that the effect of HSR is limited to the mega cities while for the rest of the sample there is no significant effect. Jiao et al. (2020) document that HSR network should not be considered as an isolate network. The HSR network facilitates accessibility to the general rail network and has an uneven impact of the growth of Chinese cities. Finally, Guo et al. (2021) provides evidence that cities with higher degree and closeness centrality in the HSR network tend to have less industrial water pollution. They rely on the reallocation of a high-skilled labour force as a mechanism which leads to the agglomeration of skill-intensive and clean industries in cities with higher HSR centrality.

Beyond China, Heuermann and Schmieder (2019) explore the expansion of HSR in Germany and find that a reduction in travel time boosts the number of commuters between regions. They argue that this effect is mainly driven by workers changing jobs to smaller cities while keeping their place of residence in larger ones. In addition, Ahlfeldt and Fed- dersen (2018) analyze the economic impact of the German HSR connecting Cologne and Frankfurt on the GDP of the three counties with intermediate stops.

More close to our research are the papers of Lin et al. (2015), Tamura (2017), Inoue and Nakajima (2017), Dong et al. (2020), Cui et al. (2020), Gao and Zheng (2020), Huang and

Wang (2020) and Komikado et al. (2021). Tamura (2017), Inoue and Nakajima (2017) and Komikado et al. (2021) investigate the effect of HSR on innovation activities in Japan. The first two papers consider the opening of the Hokuriku Shinkansen line in 1997 as an ex- ogenous shock. Tamura (2017) studies the changes in citation distances before and after the opening of the Hokuriku Shinkansen. The author finds that this HSR line facilitatesthe spreading of innovation activity in regions distant from the metropolitan area. Inoue and Nakajima (2017) use the opening of the same line to examine the effect of HSR onthe number of patents made in collaborations. They argue that there is a significant in- crease in the innovative activities of the establishments near the stations of the rail in both quantity and quality (citation data). Komikado et al. (2021) find that the existence of HSR stations has a positive and statistically significant association with knowledge productiv-ity, measured as annual patent applications divided by employees in each prefecture. Linet al. (2015) explore the effect of the introduction of HSR technology in China on the local spillovers of foreign technology transfer. They find that even though the railway firms gain the direct benefits, the diffusion process did not end there. Their analysis reveal significant spillovers to nearby firms and not only in terms of more patents, but also as higher pro- ductivity and revenue growth. Gao and Zheng (2020) combine data from three waves of innovation surveys on manufacturing firms with data on the rollout of HSR in China andfind that HSR connection boosts the innovation activity of firms. In addition, they indicate that HSR access induces the change from process innovation to product innovation. Cuiet al. (2020) explore the effect of HSR on patent applications filed by all universities and colleges in China. They apply a DID approach and find that HSR enhances patent collabo- rations across cities. In addition, they report evidence regarding the collaborations among firms and universities and intracity and intercity collaborations. Huang and Wang (2020) explore the effect of HSR network on green innovation. The authors document that HSR network fosters green innovation in general. They find also that the mechanism behind their results is the mobility of innovative factors. Finally, Dong et al. (2020) use the ex- pansion of HSR and research paper publications and citations and argue that co-authors' productivity increases when secondary cities are connected by bullet trains to China's ma- jor cities. The authors document that more new co-author collaborations take place when their cities connect to the HSR network.

We build on this literature and contribute to it in several ways. Regarding the benchmark analysis, we rely on a recent, unique dataset of disambiguated Chinese inventors (Yin et al., 2020) which covers the entire universe of patent applications at the China National Intellectual Property Administration database (CNIPA, 2017 edition). Furthermore, we

propose a novel identification strategy, which relies on couriers' stations during the Ming dynasty (1403–1644) to create exogenous variation for our main independent variable. Previous studies use either the transportation network of 1961-1962 as an instrument (Zheng and Kahn, 2013; Baum-Snow et al., 2017; Dong et al., 2020), or a difference-in-difference (DID) approach (Lin, 2017; Qin, 2017; Dong, 2018; Gao et al., 2018; Gao and Zheng, 2020), among other approaches (Hsiao et al., 2012; Ke et al., 2017; Faber, 2014; Dong, 2018; Gao et al., 2018; Banerjee et al., 2020).

We also investigate knowledge diffusion (which ultimately affects city innovation), us- ing the framework of the branching literature. To our knowledge, very few studies ad- dress the issue of diversification thanks to the connections to the outside world. Baharet al. (2020) study the effect of immigrant inventors on the technological advantage of nations, finding that countries tend to diversify in technologies brought in by migrant in- ventors originating in their countries of origin. In our case, we explore that cities are more likely to develop comparative advantage in a new technology after increasing its connec- tivity, via HSR network, with other cities already specialized in that technology. Closeto our mechanism is a very recent paper of Gao et al. (2021). The authors explore the spillovers across industries and regions in China's on regional economic diversification at the province level. They use HSR network as an instrument and they report that theirmain variables of interest, relatedness and the number of neighbouring regions facilitate the diffusion of economic capabilities. They find that the two spillover channels behave as substitutes, meaning that the marginal contribution of one channel (related industries or nearby regions) is reduced when the other is sufficiently active.

Finally, we argue that the reduction of transportation costs promotes diffusion linkages across cities like patent co-applications and mobility of the inventors. In order to understand whether these mechanisms operate, we test the effect of the expansion of HSR network on cross-city patent co-applications and inventors' mobility. We find that HSR network contributes to the specialization of cities by facilitating these diffusion linkages but it has not only limited to this.

The rest of the paper is organized as follows: Section 2 revises the rollout of HSR network in case of China. Section 3 presents the data. Section 4 shows the empirical strategy. Section 5 presents the main results, and section 6 concludes.

2 The rollout of the HSR Network in China

China has the longest HSR network in the world. China's HSR network is designed for speeds of 250–350 km/h (155–217 mph) and it was developed rapidly over the past 15 years with substantial funding from the Chinese government. The first HSR was proposed in 2004. The purpose of the network was to connect the major cities by using four horizontal and four vertical corridors. Except for the Hangzhou–Shenzhen corridor, all the new lines were built in parallel to existing conventional lines (Lawrence et al., 2019). However, in 2016 the plan was revised and updated to eight horizontal and vertical corridors to connect the medium size cities as well. Another aspect of China's HSR network is that it complements existing transportation networks (Lin, 2017). The vertical and horizontal HSR lines provide the basic network which is supplemented with regional and intercity railway lines (Lawrence et al., 2019). The final aim of HSR is to connect all the cities of 0.5 million inhabitants or more within one to four hours to a megacity.

According to the Ministry of Railway (MOR) the design of the network is based on the economic growth, population and resource distribution, national security, environmental concerns and social stability of each region (Lawrence et al., 2019). Figure 1 presents the expansion of HSR network from 2007 to 2015.

3 Empirical Strategy

We start our analysis by estimating our main model to analyse the role of HSR on the innovation performance of the Chinese cities:

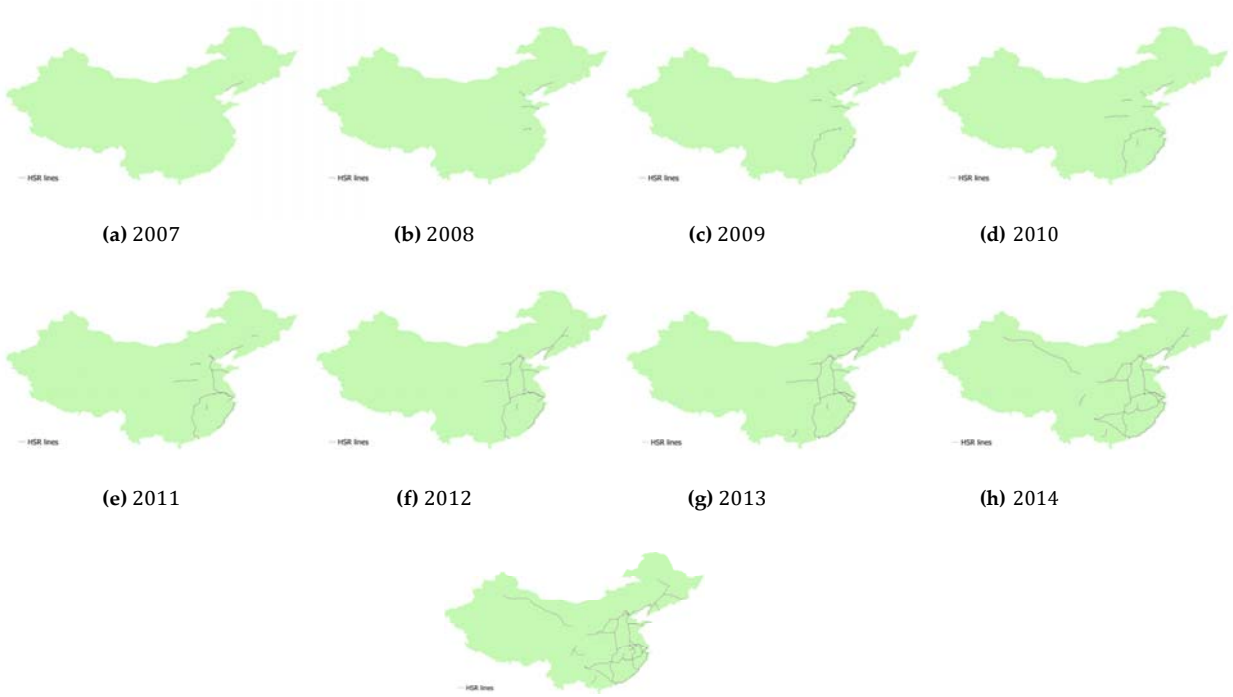
$$Y_{it} = \alpha_0 + \beta HSR_{it-1} + \varsigma_i + d_t + zX_{it} + e_{it} \quad (1)$$

where, Y_{it} is the number of patents per capita, in line with the paper of Andersson et al. (2020), in the city i, in time period t. The main variable of interest is HSR_{it-1}, which contains the straight line distance of any city most populated point to the closest HSR station as computed in the previous period.[1] We include city fixed effects, ς_i and year fixed effects, d_t. In addition, we control for the average night light activity, the average crop land areaand the average grazing area, X_{it}.[2] We estimate by OLS and cluster the standard errors atthe city level. We transform our main independent variable using the inverse hyperbolic

[1] A one-lag structure for the effect of HSR on innovation is intuitive because the stations constructed near the end of the calendar year are likely to affect innovation outcomes only in the following year (Melander, 2020).

[2] The average values of these variables are the sum divided by the number of cells in each city.

Figure 1: Expansion of high speed rail network



(a) 2007 **(b)** 2008 **(c)** 2009 **(d)** 2010

(e) 2011 **(f)** 2012 **(g)** 2013 **(h)** 2014

(i) 2015

Notes: This Figure presents the expansion of high speed rail network from 2007 until 2015. We include only lines with an average speed 250 kilometers per hour and more. Source: Li (2016) and WorldMap (2011).

sine transformation of the parameters (Bellemare and Wichman, 2020).

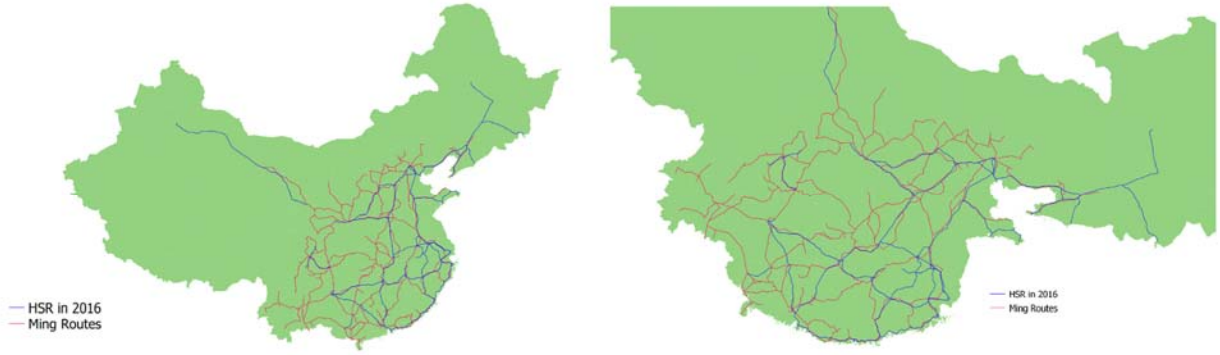
However, there is a large literature rising concerns regarding potential endogeneity issues. The most common drawback is that the placement of the HSR network could be endogenous because HSR stations are not randomly established. As a result, transporta- tion investments may be correlate with outcomes of interest (Andersson et al., 2020) or assigned through the existing political process (Redding and Turner, 2015). For these rea- sons, we adopt an identification strategy based on instrumental variables.

3.1 Identification Strategy: Historical Couriers' Routes and Stations

We rely on couriers' routes and their stations during the Ming dynasty (1368-1644) to create exogenous variation for our analysis. Previous studies use either network infrastructures (Baum-Snow et al., 2017, 2020) or straight line instruments (Banerjee et al., 2020) or DID approaches (Dong, 2018). We have several reasons to believe that our approach

contributes to the existing ones. First, we do not base our identification strategy on an old infrastructure network. An old infrastructure network could correlate with variables that predicts economic outcomes of interest (Baum-Snow et al., 2017). We argue that, in our case, the routes represent the least cost paths which couriers used during Ming dynastyto deliver their messages. Second, we do not use recent land features like in the paper of Andersson et al. (2020) to construct our least cost routes. Figure 2 presents the routes during Ming dynasty with red lines while with the blue lines are the lines of HSR network. We argue that the historical couriers' routes as least cost paths could be strong predictors of the recent HSR lines. We believe that Ming dynasty is the appropriate time period because, according to Ma et al. (2016), the efficient distribution system of couriers reached at its peak.

Figure 2: HSR lines in 2016 and Ming routes



(a) panoramic view of China **(b)** mainland of China

Notes: Figure 2 presents the high speed rail lines in 2015 and Ming routes. Source: Li (2016) and WorldMap (2011).

We provide evidence that these routes could assemble a least cost network. Twitchett and Mote (1998) describe in detail the organization of the courier services and provide illustrative examples. They argue that there were 1936 operational stations established at a distance of 35 to 40 km, one from the other. The purpose of the courier routes was not to connect every county, but to link the provincial and prefectural capitals. They were defined as the major routes of the Ming dynasty. Furthermore, we have strong reasons to believe that the couriers used the fastest routes. According to Twitchett and Mote (1998) there was a penalty for the couriers for exceeding the time limits to deliver the mail.[3]

[3] For exceeding the time limit by a day, a courier was liable to a beating of twenty strokes, plus an addi-

We compute in QGIS the straight line distance from any most populated point to the closest courier station. Since the instrument is static, we multiple it with time dummies (Andersson et al., 2020; Melander, 2020). We expect a positive relation between the instrument and the HSR stations. A city which is far away from the courier stations during Ming dynasty should also be far away from the HSR stations in recent time period.

Next, we form our first stage equation:

$$HSR_{it-1} = \alpha_0 + \sum_y k_y(Distance\ to\ Courier\ Station)_i + \varsigma_i + d_t + \mathbf{z}X_{it} + e_{it} \quad (2)$$

where κ_y is a variable which takes the value 1 if the year is equal to y = 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016 and 0 otherwise. Distance to courier station has been transformed using the inverse hyperbolic sine transformation. Consequently, in a second stage, our main equation is:

$$Y_{it} = \alpha_0 + \beta\hat{HSR}_{it-1} + \varsigma_i + d_t + \mathbf{z}X_{it} + s_{it} \quad (3)$$

where $\hat{HSR}_{it-1}$ is the fitted value of HSR according to equation 2 and Y_{it} is the number of patents per capita. The rest of the control variables are the same as in the equation 1.

3.2 The role of HSR on the diffusion of knowledge across cities

3.2.1 HSR and inventors' mobility and co-patenting activities

In this section, we particularly look at the effect on cross-city patent co-applications and inventors' mobility. Breschi and Lissoni (2009) explore the effect of mobile inventors and networks of inventors to the diffusion of knowledge across firms and within cities or states. They find solid evidence that the most important reason why geography matters in constraining the diffusion of knowledge is that limits the mobility of the researchers. As a result, their co-invention network is also localized. D'Este et al. (2013) study the effect of geographical proximity on research collaborations between universities and industry.

tional stroke for every three days beyond that, to a maximum of sixty (Twitchett and Mote, 1998). Next, we provide a story as anecdotal evidence from the book of Twitchett and Mote (1998): "On 23 March, the Hangchow prefectural government assigned Ch'oe's party a different escort and issued them with a document empowering them to travel by the courier service. His escort was given an arrival deadline of 11 May, with threat of punishment should he fail to meet it. Ch'oe was told informally that the journey from Hangchow to Peking would take about forty days, though the deadline gave them forty-seven days in which to get to the capital." Using the couriers' routes, they arrived in Peking two days before their travel permit expired.

They state that geographical proximity makes universities and industry research partnerships more likely. Breschi and Lenzi (2016) they find that co-invention networks is crucial for inventive productivity in a sample of US cities. They argue that the co-invention net- work matters when there is high social proximity between the members in the city andlocal cliques of co-inventors in which interaction is dense. A developed co-invention net- work facilitates the interactions among inventors. The inventors built their own knowledge by interacting with and learning from others. Interactions with better inventors are very strongly correlated with higher subsequent productivity (Akcigit et al., 2018; Tubiana et al., 2020). Finally, the mobility of the inventors enhances patent collaborations by helping them to overcome localization constraints (Miguelez, 2019).

We introduce a gravity equation such as:

$$Know_{odt} = \alpha_0 + \beta Cost_{odt-1} + \varsigma_{od} + d_{ot} + z_{dt} + e_{odt} \quad (4)$$

where $Know_{odt}$ could be either the number of co-applications or the flow of inventors across two cities in year t. o is the symbol for origin city and d the symbol for destination city. $Cost_{odt-1}$ is the transportation cost based on HSR lines across two cities as it was computed in $t-1$ period. We include, ς_{od}, origin - destination fixed effects, d_{ot}, origin - year fixed effects and z_{dt}, destination - year fixed effects. We double cluster the standard errors at origin and destination city. We use the OLS model of Correia (2015) to estimate the gravity equation.

Admittedly, it would be ideal to measure collaborations among inventors across different regions. However, CNIPA database does not contain information of inventor's addresses but only the address of the first applicant (Yin et al., 2020). Assuming inventors from different organizations would involve in co-application, we adopt cross-regional co-application to proxy for regional collaboration. After filling in the address and thus geocoding data of non-first applicants with their corresponding information provided in other patents, we identify 490,314 co-application cases in which 260,030 (about 53.03%) are cross-city collaborations until 2016.

In contrast with the co-application data, the data about the mobility of the inventors is available until 2014. Moreover, due to the similar data limitations mentioned above, we identify mobile inventors based on the change of organizations. This means that using patent data, to be identified as a mover an inventor needs to have at least two different ap-

plicants subsequently throughout his/her patenting history (Marx et al., 2009).[4] In addi- tion, to minimize the possibility of generating fake movers, we filter out moving inventors with very conservative and strict selection criteria in stemming and matching applicant's names. As result, among 2.59 million inventors with more than 2 patents, we filtered 20,532(about 0.79%) inter-organizational mobility cases who moved before 2015.

3.2.2 HSR and technological diversification of cities

We estimate the following equation using the OLS model of Correia (2015), with high dimensional fixed effects:

$$entry_{ict} = \alpha_0 + \beta Cost_{ict-1} + \theta RelDen_{ict-1} + \varsigma_{ic} + d_{ct} + h_{it} + e_{ict} \quad (5)$$

where $entry_{ict}$ is a binary variable which switches to 1 if the city i starts to specialize in a given technology c in year t. $Cost_{ict-1}$ is the median transportation cost to the cities that are specialized in a given technological field. $RelDen_{ict-1}$ is the relatedness density of the city. We include, city-year fixed effects, h_{it}, ipc-year fixed effects, d_{ct} and city-ipc, ς_{ic}, fixed effects. We also scale the main independent variable to deal with small values (Bellemare and Wichman, 2020) and we transform it using the IHS transformation. We cluster the standard errors at the city level.

Finally, we also include in equation 5 the flows of inventors and the co-applications[5] from the cities that already have a RTA larger or equal to 1 to the rest of the cities in the sample. Our last equation is:

$$entry_{ict} = \alpha_0 + \beta Cost_{ict-1} + \theta RelDen_{ict-1} + \gamma Inv_{ict-1} + \mu Co_{ict-1} + \varsigma_{ic} + d_{ct} + h_{it} + e_{ict} \quad (6)$$

where γInv_{ict-1} and μCo_{ict-1} are the inflow of inventors from the cities that are already specialized in a given technology and co-applications with the cities that are already specialized in a given technology, respectively.

[4] Therefore, the identified mobile cases ends in 2015.

[5] We have add 1 and multiple by 100 the flows of the inventors and the co-applications to avoid zeros and small values in the data that could cause distortions before we use the inverse hyperbolic transformation.

4 Data

4.1 Patent Data

We rely on a unique dataset of disambiguate Chinese inventors. Our source is the China National Intellectual Property Administration database (CNIPA, 2017 edition). CNIPA was founded in 1980 as the Patent Office of the People's Republic of China. As Yin et al. (2020) indicate, inventor level research in East Asia suffers from the common name problem, i.e., many Chinese share the same name. For example, a search of the name Zhang Wei hits up to 9,680 patents and 61,037 papers in Wanfang Chinese DBLP database, which clearly indicates a problem of homonymy. Our patent database does not suffer from these drawbacks. It relies on a systematic disambiguation process carried out by means of machine learning techniques, using data from 1985 to 2016. It contains the full universe of Chinese patents, 4,967,900.

The assignment of inventors to cities is based on reverse geocoding using CNIPA's addresses and Baidu Map's API (Yin et al., 2020). CNIPA collects only the first applicant's address. We allocate the inventors to their cities based on the latitude and longitude of their address. Our shapefile comes from the National Geomatics Center of China and contains 349 cities for mainland China. We aggregate patent data at the city level to have a measure of city innovation.

4.1.1 Technological specialization

We look at knowledge diffusion, building on the branching literature mentioned above. We classify patents in technological classes according to the 4 digits IPC codes. We restrict our analysis to technological classes that appear in every year. Our final database includes 599 technological classes in total. We rely on the relative technological advantage index to define our entry variable and also to construct our independent variables, transportation cost, inflow of inventors, co-applications and relatedness density, for the model of knowledge diffusion. We define the relative technological advantage index for each city as:

$$RTA_{ict} = \frac{pat_{ict} / \sum_c pat_{it}}{\sum_i pat_{ct} / \sum_c \sum_i pat_t} \tag{7}$$

where pat_{ict} is the number of patents that city i produced in technology c in time t. This index relies on Soete (1987) and it is similar to the Revealed Comparative Advantage(RCA) index by Balassa (1965). Based on this index we define our main independent and

dependent variables as explained next.

We define the variable "entry" as a binary indicator which switches to 1 if the city in a given year starts to specialize in a specific technology. It becomes 1 only when the city *i* in a year *t* has an RTA larger or equal than 1 while in the period $t-1$ it was an RTA less than 1. If the city preserves an RTA larger or equal than 1 the following years it becomes missing value since a city cannot enter again, by definition, before exiting. As long as a city has an RTA less than 1 and also the RTA in the period $t-1$ was less then 1 it preserves the zero value.

Our main independent variable, for a given city *i* in a given year *t* and for a given technology *c* is the median transportation cost to the cities that are specialized in the technology *c* (or equally cities that have an RTA larger or equal to 1). Thus, the computation of the cost for every city, year and technological class depends also on the cities that are specialized. Furthermore, it could be possible to observe an increase in our cost variable because there is an increase in the number of cities that are specialized in a given technology *c* in the year $t+1$ comparing to the year *t*. This is the reason why we use the median cost in our benchmark analysis. As a robustness test, we provide the results using the total cost in the Appendix. The intuition behind the use of the median transportation cost is that it does not get affected from extreme values in the sample as total cost could possibly be affected.

Following the literature on the technological specialization, we compute the relatedness density variable. First, we created the co-occurrence matrix and we normalize it using the association probability measure, proposed by van Eck and Waltman (2009). We compute one co-occurrence matrix for each of the time periods under consideration. We rely on this matrix to develop the relatedness density index for each set region-technology that indicates how close a technology is to the existing technological base of a city. The technological base of a city is considered to consist of those technologies in which a city has developed a specialization, measured as the Revealed Technological Advantage (RTA). We use the Econ Package in R to compute the relatedness density index (Balland, 2017). With this variable we measure to what extent the technological base of a city is related to technology i. The higher the degress of relatedness, the easier we expect city i to specialize in this technology.

4.2 Rail Data

Our paper makes use of three different sources to construct the HSR network. We extract data about HSR stations and lines of China's High Speed Railway System from Li (2016) and WorldMap (2011) webpage. We explore the shapefiles of both sources to find the most precise and accurate placement of lines and stations. In addition, we extract data about the speed of each line in two different time periods.[6] Our final source of data is the official website of the National Railway Administration of the People Republic of China which contains the opening years of the HSR stations (www.12306.cn).

As shown in Figure 1, we consider only the lines with an average speed of more than 250 kilometers per hour - in line with the definition of Dong et al. (2020). Our setting allows us to compute in QGIS the straight line distance in meters from the most populous points of every city to the closest HSR station.[7] We prefer this measure over a binary variable in terms of higher variation since the HSR network is accessible only through the stations, in contrast with a road network. We believe that our variable captures better this effect than a binary indicator. Huang and Wang (2020) apply a similar approach based on the spherical distance to the closest HSR station. To identify the most populous points we use the raster file of population in 2007, one year before the arrival of HSR network, from HYDE (2020) database. We transform the pixels into points and we choose for every city the pixel with the highest population value. An alternative approach would be to select the city centroids instead of the most populous points but we believe that our setting is more realistic.

4.2.1 Computation of the transportation cost

The HSR network has been responsible for an enormous reduction in transportation costs in China. In this section, we shed light of how this reduction affected the specialization of cities.We start our analysis by computing the accessibility of each city based on the expansion of HSR network in QGIS.

1. We divide China in $0.09^o \times 0.09^o$ grids. One grid is approximately 80 square kilometers.

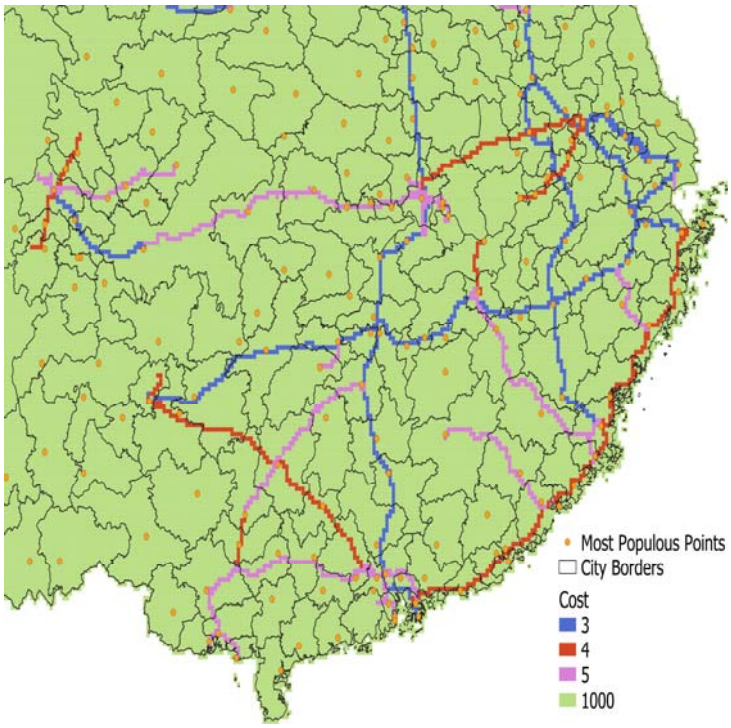
2. We use the average travel speed of each line in order to allocate costs to our grids. We have 3 categories of lines according to their travel speed. Their speed could be 200, 250

[6] WorldMap (2011) is until 2011 while Li (2016) is updated until 2016.
[7] We have information about the latitude and the longitude of every HSR station.

or 300 kilometers per hour on average. Then, we create our cost values using the reverse of the average travel speed. A line with an average speed of 300 kilometers per hour is assigned a value of 3, for a line with 250 kilometers per hour the value of 4 and for a line with 200 kilometers per hour the value of 5.[8]

3. Then, we allocate to each grid a cost value of 3, 4 or 5 depending if it is intersected with a HSR line and the category of the line. If no line is intersected with a grid then thegrid takes the value of 1000.[9] Figure 3 presents an example for the year 2016.

Figure 3: Cost allocation based on the HSR lines



Notes: This graph presents the cost allocation to the grids based on the HSR line network in 2015. The orange dots are the most populous points and the black lines are the borders of the cities. The blue grids intersect with a line that have an average speed of 300 kilometers per hour, the red girds intersect with a line that has an average speed of 250 kilometers per hour and finally, the pink grids intersect with a line that has an average speed of 200 kilometers per hour. Authors' computations.

[8] We compute these values for each grid by simple doing (1/average travel speed of HSR line)*1000.

[9] We compute the values by using the reverse of the average travel speed. Our cost values are between 0 and 1, before we multiple them by 1000. We assign the value 1 to the grids that not crossed by a HSR line. Next, we multiple the value 1 by 1000. The intuition to assign a such high value is that these grids can onlybe crossed by walking. Even, if there is an alternative transportation network, like roads, since it does not change dramatically over our study period, from 2007-2016, its effect should be captured by the time fixed effects. Over this time window, the government's focus was mainly the expansion of the HSR network.

4. Next, we create our cost raster files, one for each year.

5. We apply Dijkstra (1959) algorithm to compute the least cost paths for a given city most populated point *i* to all the other cities' most populated points except *i*. It is a typical minimisation problem based on the least cost surface which selects the optimal route. The result is 123,201 least cost paths for every year which can be summarized in the following cost matrix:

$$C_t = \begin{bmatrix} \text{Cost}_{11}^{-\theta} & \text{Cost}_{12}^{-\theta} & \dots & \text{Cost}_{1n}^{-\theta} \\ \text{Cost}_{21}^{-\theta} & \text{Cost}_{22}^{-\theta} & \dots & \text{Cost}_{2n}^{-\theta} \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cost}_{n1}^{-\theta} & \text{Cost}_{n2}^{-\theta} & \dots & \text{Cost}_{nn}^{-\theta} \end{bmatrix}_t$$

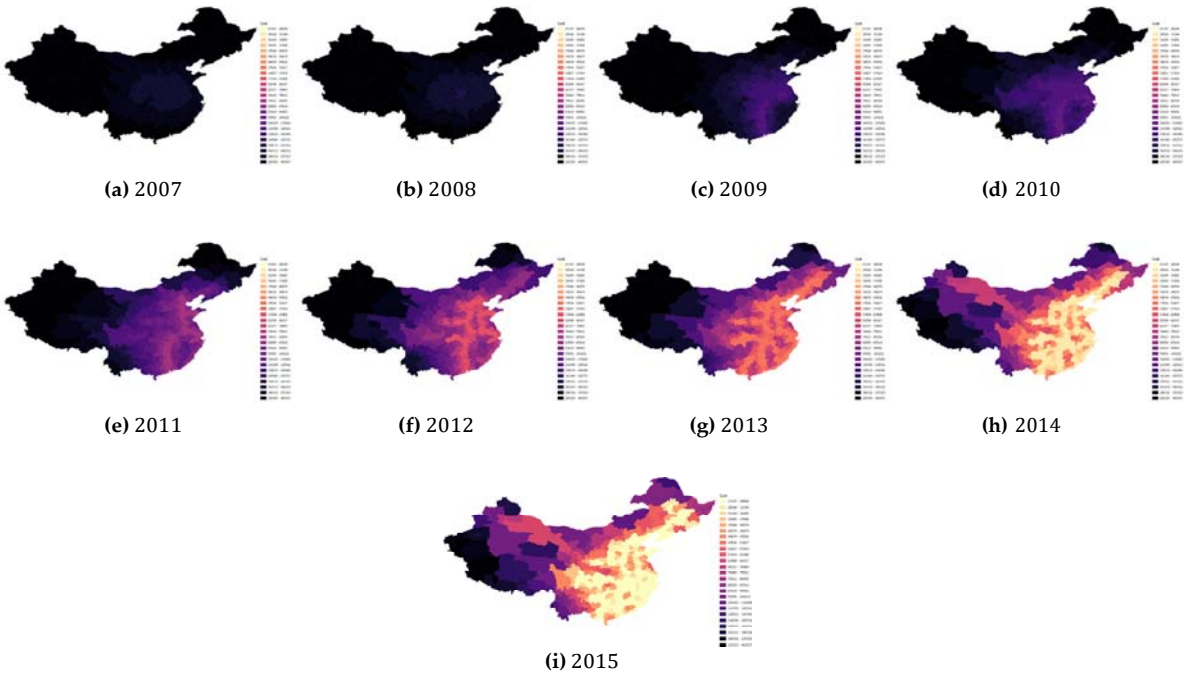
6. Finally, we compare the costs between city pairs and between two different consecutive years (t and $t+1$) and we assign always in year t the lower of the two costs (Perlman, 2015).

We then aggregate the cost for a given city *i* to move to all the other cities except *i* and the result of our exercise appears in Figure 4. Along the period considered, it is clearly observed a reduction in transportation cost due to the expansion of HSR network. This is a typical approach used in the literature to compute accessibility and to define market access (Donaldson and Hornbeck, 2016; Hornbeck and Rotemberg, 2019). Next, we use this computed transportation cost to explore how the connectivity to a HSR network is associated with the technological specialization of cities (Bahar et al., 2020; Balland et al., 2019).

4.3 Other data and Variables

In order to control for economic activity at the city level we rely on night light data. According to the previous literature, night light data is a decent proxy for economic activity (Henderson et al., 2011; Mellander et al., 2015). In addition, Kulkarni et al. (2012) report evidence that for only a very low count of prefectures (between 5 to 10%) the night lights are not a good proxy for determining local GDP in case of China. We make use of the har-

Figure 4: Transportation costs



(a) 2007 **(b)** 2008 **(c)** 2009 **(d)** 2010

(e) 2011 **(f)** 2012 **(g)** 2013 **(h)** 2014

(i) 2015

Notes: Following method in section 4.2.1 this graph presents the reduction in transportation costs. Authors' computations.

monized global nighttime light dataset in the period 1992-2018 given in Li et al. (2020).

We extract a large set of control variables from HYDE (2020) database (Goldewijk et al., 2010, 2011).[10] HYDE (2020) provides (gridded) time series of population and land use for the last 12,000 years. We use raster files to determine the population, cropland and grazing area at the city level.

Finally, we control for the access to airports since recent literature finds a positive association between air connectivity and patenting (Wong, 2019). We make use of the Civil Aviation Administration of China website (CAAC, n.d.) to obtain data regarding the number of passengers. All these data comes at the airport level so that we use reverse geocod- ing in the names of the airports and aggregate the passenger data at the city level. Table1 separates the variables, we use, by estimation model and presents the summary statistics.

[10] We use the baseline estimates of the version 3.2.

Table 1: Summary statistics

Variable	Mean	Std. Dev.	Min.	Max.	N
Model of innovation performance					
Patents per 10,000 People	2.1344	4.8374	0	54.1197	3141
Distance to HSR stations	670647.8034	894601.0487	864.9210	4857224.5	3141
Distance to Ming stations	2271453.5489	4921132.4595	322.2391	24786600	3141
Average night light area	7.1898	8.1819	0.0014	58.3364	3141
Average crop area	18.7441	12.6208	0.0018	44.6337	3141
Average grazing area	19.1303	13.3489	0.3243	61.6345	3141
Air transportation (passengers)	1929347.2098	8063795.7783	0	107543629	3141
Model of knowledge diffusion					
Entry	0.0883	0.2837	0	1	1631595
Median transportation cost X RTA	326039.5001	317719.5295	117.338	1822684.5	1631595
Patent co-applications X RTA	14.1078	93.3737	0	6644	1631595
Inflow of inventors X RTA	0.4797	1.9777	0	60	1631595
Relatedness density	5.4251	16.745	0	100	1631595
Mobility - Gravity Model					
Inventors	0.0074	0.1404	0	16	845292
Total transportation cost	475507.0006	383637.1529	58.669	2041284.625	845292
Patent Co-applications - Gravity Model					
Patent collaborations	0.2172	5.4278	0	1248	546534
Total transportation cost	407232.361	374537.709	25	2041284.625	546534

Notes: Summary statistics for all the variables (Model of innovation performance: 349 cities from 2008-2016. Model of knowledge diffusion: 348 cities from 2008-2016. Mobility - Gravity model: 348 cities from 2008-2014. Patent collaborations - Gravity model: 348 cities from 2008-2016.). **Model of innovation performance:** The number of patents per capita is the number of patent applications for every city divided by the population of the city. Distance to HSR Station 250 is the computed straight line distance (in meters)from the most populated point of a city to the closest high speed rail station. The rest of the control variables are the computed average night light area for every city, the cropland area for every city, the average land used for grazing for every city and as air transportation, we control for the number of passengers. Instrument: Distance to Ming stations is the distance in meters from the most populous point of a city to the closest courier station during Ming dynasty. **Model of knowledge diffusion:** The entry variable is a binary indicator which switches to 1 if the city in a given year starts to specialize in a specific technology. Median transportation cost is the median computed transportation cost using the HSR to the cities that are specialized in a given technology. Patent co-applications are the patent co-applications of every city with the cities that are specialized in a given technology. Inflow of inventors are the inventors that move from one city that specialize in a given technology to the other cities. The relatedness density index foreach set region-technology indicates how close a technology is to the existing technological base of a city. The technological baseof a city is considered to consist of those technologies in which a city has developed a specialization, measured as the Revealed Technological Advantage (RTA). **Mobility - Gravity model:** The inventors are the labor flows of the inventors that move from one city to the other. The total transportation cost is the total computed transportation cost using the HSR between pair of cities. **Patent collaborations - Gravity model:** The patent co-applications are the patent co-applications between pair of cities. The total transportation cost is the total computed transportation cost using the HSR between pair of cities.

5 Results

5.1 The role of HSR on innovation performance

Table 2 contains the first stage estimates. The effect of the instrument on the distance to the high speed station solely comes from cross-sectional variation. Based on that, the positive sign means that for a given city if it is close to a courier station during Ming dynasty is also close to a HSR station in a more recent time period. Both variables have the same initial unit of analysis, meters. We standardize all our variables of interest.

Table 2: First stage: Distance to courier stations and distance to HSR stations

Dep. var. =	IHS Distance to Stations		
	(1)	(2)	(3)
IHS Distance to Courier Station X 2009	0.0693***	0.0680***	0.0647***
	[0.0099]	[0.0098]	[0.0091]
IHS Distance to Courier Station X 2010	0.1388***	0.1373***	0.1323***
	[0.0141]	[0.0141]	[0.0134]
IHS Distance to Courier Station X 2011	0.1738***	0.1742***	0.1670***
	[0.0155]	[0.0156]	[0.0151]
IHS Distance to Courier Station X 2012	0.1607***	0.1616***	0.1585***
	[0.0162]	[0.0162]	[0.0157]
IHS Distance to Courier Station X 2013	0.1639***	0.1651***	0.1654***
	[0.0163]	[0.0163]	[0.0157]
IHS Distance to Courier Station X 2014	0.1792***	0.1804***	0.1826***
	[0.0172]	[0.0172]	[0.0166]
IHS Distance to Courier Station X 2015	0.1487***	0.1500***	0.1525***
	[0.0206]	[0.0208]	[0.0212]
IHS Distance to Courier Station X 2016	0.1334***	0.1346***	0.1382***
	[0.0207]	[0.0210]	[0.0214]
IHS Average Night Light	-0.4592***	-0.4314***	-0.4077***
	[0.1033]	[0.1035]	[0.0912]
Sample Size	3141	3105	2988
City FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Sample	Full	Second and Third	Third

Notes: First stage regression results based on equation 2. The dependent variable is the straight line distance in meters from the most populous point of a city to the closest high speed railway station. Distance to a courier station is the straight line distance in meters from the most populous point of a city to the closest courier station. Average night light data controls for the economic activity of a city. We control for the average cropland area and the average land used for grazing. The variables distance to a HSR station and distance to a courier station have been transformed using the inverse hyperbolic sine transformation. Clustered standard errors at the city level are reported in the brackets.

Next, we move to our benchmark results in Table 3. Panel A contains the OLS results and panel B the IV. Both OLS and IV coefficients coefficients for distance to HSR are significant and negative meaning that for a given city a reduction in the distance to the closest station is associated with an increase in the number of patents per capita. More specifically,

one standard deviation decrease in the distance to a HSR station results in an increase by 0.246% in the patents per capita, panel B column 1. In the second column, we remove the first tier cities according to the definition of Fang et al. (2015). We see that the coefficient of the IV estimate is now getting reduced. In the final column we keep only the third tier cities. In column 2 and 3 we observe a reduction in the IV coefficients of distance to HSR station and become significant at 5%, compared to column 1 which is at 1%. The exclusion of the big urban centers affects the magnitude of the IV estimates. However, the fact that the coefficients are still significant at 5% shows that the diffusion process is partially driven also from the second and third tier cities. Both control variables in this regression, night light activity and passengers by air, have a positive effect on innovation activity.

Table 3: Main Results: HSR and Innovation

Dep. var. =	Patents per capita		
	(1)	(2)	(3)
Panel A	**OLS**	**OLS**	**OLS**
IHS Distance to Stations	-0.1183***	-0.1201***	-0.1225***
	[0.0343]	[0.0342]	[0.0364]
Average Night Light	0.6184***	0.6557***	0.6387***
	[0.1813]	[0.1787]	[0.1866]
Flight Passengers	0.5609***	0.6008***	0.4164**
	[0.1696]	[0.1487]	[0.1859]
R-squared	0.83	0.77	0.75
Panel B	**IV**	**IV**	**IV**
IHS Distance to Stations	-0.2461***	-0.2176**	-0.2144**
	[0.0868]	[0.0848]	[0.0851]
Average Night Light	0.5454***	0.6023***	0.5906***
	[0.1852]	[0.1779]	[0.1863]
Flight Passengers	0.5365***	0.5749***	0.3940**
	[0.1722]	[0.1527]	[0.1887]
First-Stage F-stat	18.52	18.47	18.97
Sample Size	3141	3105	2988
City FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Sample	Full	Second and Third	Third

Notes: Main results. The dependent variable is the number of patents divided by population. Distance to high speed rail stations contains the straight line distance in meters from the most populous point of a city to the closest high speed railway station. Average night light data controls for the economic activity of acity. We control for the average cropland area and the average land used for grazing of a city. The variable distance to a HSR station has been transformed using the inverse hyperbolic sine transformation. Panel A contains the OLS results based on equation 1 and panel B the IV estimates based on equation 3. Clustered standard errors at the city level are reported in the brackets.

According to Dong et al. (2020) an explanation for the larger IV estimates is that the Chi-

nese government intentionally planned some HSR stations to lagging areas to help them grow. This could explain the downward bias in the OLS estimates.

Finally, our results are not affected even when we use the stations with access to a line that has an average speed of 200 kilometers or more. As a robustness test we repeat the same regressions and we include in our analysis these stations in Table 7 in the Appendix.

5.2 The role of HSR on knowledge diffusion

Table 4 presents the regression results of transportation cost on co-applications as in equation 4. According to column 1, we find that one standard deviation decrease in the transportation cost is associated with an approximately 0.046% increase in patent co-applications across cities. In column 2, we split our focal variable in ranges of 250 km.[11] We see that the cities which are close in terms of distance are the ones that are most benefited in terms of patent co-applications. However, the next two categories which are benefited the most are the cities that are most far away with a coefficient 0.1697 and 0.1541.

Next, we move on to the Table 5 and show the regression results of transportation costs on inventors' mobility. According to the first column, the coefficient is negative but not significant. Similar to the Table 4, we regress in the 2nd column the different categories on the mobility of inventors. The evidence of the 2nd column is ambiguous. The coefficients of the first two categories are negative meaning that for the cities which have a straight line distance, between them, of less than 500 km a reduction in transportation increases the mobility of the inventors. Then, for the cities in the middle, there is no effect of transportation cost on the mobility of inventors. In the last category the coefficient is positive and significant meaning that a reduction in transportation cost reduces the mobility of inventors. One plausible explanation for the positive sign could be that the travelling cost via the HSR for the citizens of these pairs of cities is prohibitive so instead of commuting they decide to change their employer.

Table 6 summarizes the results of the technological specialization model model. The median transportation cost has a negative and strong effect on the probability of entry. One standard deviation decrease in the transportation cost of a city i for a given technology c to the other cities that are specialized in this technology c increases on average the probability of the city i to specialize in the technology c by 0.0636%. In column 2 we test

[11] We chose these thresholds because the average speed of HSR network is 250 km per hour and it needs one hour to cover a distance of 250 km.

Table 4: Transportation cost and patent co-applications - OLS

Dep. var. =	Co-applications	
	(1)	(2)
IHS Transportation Cost	-0.2200**	
	[0.0993]	
Less than 250 km		-5.2621***
		[1.2108]
250-499 km		-1.7035
		[1.0521]
500-749 km		-0.8481*
		[0.4847]
750-999 km		-0.3845
		[0.5489]
1000-1249 km		-1.5054
		[1.5293]
More than 1250 km		0.6636
		[0.8525]
Sample Size	546516	546516
City Origin x Destination FE	Yes	Yes
Destination FE x Year FE	Yes	Yes
Origin FE x Year FE	Yes	Yes

Notes: Gravity model based on equation 4 estimated with OLS and standard errors clustered at the origin and destination city. The patent co-applications are the patent co-applications between each pair of cities. The total transportation cost is the total computed transportation cost using the HSR between each pair of cities.

Table 5: Transportation cost and the mobility of inventors - OLS

Dep. var. =	Mobility	
	(1)	(2)
IHS Transportation Cost	-0.0001	
	[0.0035]	
Less than 250 km		-0.5004***
		[0.1186]
250-499 km		-0.0626***
		[0.0185]
500-749 km		-0.0018
		[0.0151]
750-999 km		0.0203
		[0.0141]
1000-1249 km		0.0356**
		[0.0173]
More than 1250 km		0.0820***
		[0.0226]
Sample Size	845292	845292
City Origin x Destination FE	Yes	Yes
Destination FE x Year FE	Yes	Yes
Origin FE x Year FE	Yes	Yes

Notes: Gravity model based on equation 4 estimated with OLS and standard errors clustered at the origin and destination city. The inventors are the labor flows of the inventors that move from one city to the other. The total transportation cost is the total computed transportation cost using the HSR between each pair of cities.

the effect of patent co-applications on the probability of a city to specialize. We find that co-applications across cities reinforces the probability of a city to specialize in a given technology in which other cities are specialized. In column 3 we restrict our sample until 2014 in order to make it comparable with the model that uses the mobility of inventors as a variable of interest mobility is only available until 2014). We introduce in our model the flow of inventors from the cities that are specialized in a given technology to the rest of the cities in the sample (columns 4 and 5). Our evidence suggests that one standard deviation increase in the number of flows of inventors is associated with a 0.0115% in the probability of a city to specialize in a given technology. In column 6 of Table 6, we explore the effect of mobility and patent co-applications at the same time on the probability of a city to specialize in anovel technological field. Both of them are positive and highly significant. One standard deviation increase in the number of patent co-applications is associated with a 0.0154% increase in the probability of a city to specialize in a given technology. In the column 7, we include in the same regression all our measures that affect the probability of entry. In all the specifications the relatedness density is highly significant and positive. One standard deviation increase in the relatedness density index corresponds to a 0.0076% increase in the probability of a city to gain technological advantage in a given technology, column 7.

Table 6: Technological specialization model

Dep. var. =	Entry
	(1)
IHS Change of Median Transportation Cost	0.0021***
	[0.0007]
Relatedness Density	0.0024***
	[0.0004]
R-squared	0.31
Sample Size	1629435
City*IPC FE	Yes
City*Year FE	Yes
Year*IPC FE	Yes

Notes: The dependent variable is binary and switches to 1 if a city has a comparative advantage in a specific technological field. Median transportation cost is the median computed transportation cost using the HSR to the cities that are specialized in a given technology. Patent co-applications are the patent co-applications of every city with the cities that are specialized in a given technology. Inflow of inventors are the inventors that move from one city that specialize in a given technology to the other cities. The relatedness density index for each set region-technology indicates how close a technology is to the existing technological base of a city. The technological base of a city is considered to consist of those technologies in which a city has developed a specialization, measured as the Revealed Technological Advantage (RTA). We include city*ipc fixed effects, city*year fixed effects and year*ipc fixed effects. OLS model based on equations 5 and 6 with clustered standard errors at the city level are reported in the brackets.

6 Conclusion

This paper explores the role of the expansion of the HSR on innovation activity in China. Firstly, our analysis contributes to the growing literature about the effect of transportation networks on innovation activity (Andersson et al., 2020; Agrawal et al., 2017; Perlman, 2015; Tamura, 2017; Inoue and Nakajima, 2017; Cui et al., 2020) by using a brand new disambiguated Chinese inventors dataset (Yin et al., 2020) covering the entire universe of CNIPA patent applications. In addition, we rely on historical couriers' stations as a novel instrument, to create exogenous variation for our analysis. We report evidence that connectivity to HSR enhances the patenting activity of a city. Our results remain highly significant when we remove the large urban centers meaning that the effect on innovation is also driven by second and third tier cities. Our results are also robust to air connectivity.

Secondly, we aim at providing evidence of some of the mechanisms that may work when analyzing the role of the HSR on innovation diffusion. We focus on the impact that HSR may have on the mobility of inventors, on the collaboration across inventors in different cities as well as on the technological diversification of cities. In order to do it, we compute transportation costs for every pair of cities based on the speed of HSR lines. We provide evidence that the reduction in transportation costs fosters patent co - applications across cities. In addition, we argue that the roll-out of HSR network facilitates labor flows of inventors for the cities that are within 500 km distance. We contribute in the literature about the factors that drive the diversification of regions (Petralia et al., 2017; Balland et al., 2019; Rigby, 2015; Bahar et al., 2020) by presenting for the first time evidence that the probability of a city to specialize in a new technological field, comes as a result of the specialization patterns of the cities to which the city connects through HSR. In addition, we argue that HSR promotes diffusion linkages such as the mobility of the inventors and co-applications which enhance the probability of a city to specialize in order to rule out its impact so as to have the net impact of the reduction of costs thanks to HSR.

References

Agrawal, Ajay, Alberto Galasso, and Alexander Oettl, "Roads and innovation," *Review of Economics and Statistics*, 2017, *99* (3), 417–434.

Ahlfeldt, Gabriel M and Arne Feddersen, "From periphery to core: measuring agglomeration effects using high-speed rail," *Journal of Economic Geography*, 2018, *18* (2), 355–390.

Akcigit, Ufuk, Santiago Caicedo Soler, Ernest Miguelez, Stefanie Stantcheva, and Valerio Sterzi, "Dancing with the Stars: Innovation Through Interactions," *NBER Working Paper 24466*, 2018.

Andersson, David, Thor Berger, and Erik Prawitz, "Making a Market: Infrastructure, Integration and the Rise of Innovation," *Working Paper Series from Research Institute of Industrial Economics*, 2020, *1319.*

Bahar, Dany, Prithwiraj Choudhury, and Hillel Rapoport, "Migrant inventors and the technological advantage of nations," *Research Policy*, 2020, *49* (9), 103947.

Balassa, Bela, "Trade Liberalisation and "Revealed" Comparative Advantage," *The Manchester School*, 1965, *33* (2), 99–123.

Balland, Pierre Alexandre, "Economic Geography in R: Introduction to the EconGeo Package," *SSRN Electronic Journal*, 2017.

_ , **Cristian Jara-Figueroa, Sergio G. Petralia, Mathieu P.A. Steijn, David L. Rigby, and César A. Hidalgo**, "Complex economic activities concentrate in large cities," *Nature Human Behaviour*, 2020, *4* (3), 248–254.

_ , **Ron Boschma, Joan Crespo, and David L. Rigby**, "Smart specialization policy in the European Union: relatedness, knowledge complexity and regional diversification," *Regional Studies*, 2019, *53* (9), 1252–1268.

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

Baum-Snow, Nathaniel, J. Vernon Henderson, Matthew A. Turner, Qinghua Zhang, and Loren Brandt, "Does investment in national highways help or hurt hinterland city growth?," *Journal of Urban Economics*, 2020, *115*, 103124.

_ , **Loren Brandt, J. Vernon Henderson, Matthew A. Turner, and Qinghua Zhang**, "Roads, railroads, and decentralization of Chinese cities," *Review of Economics and Statistics*, 2017, *99* (3), 435–448.

Bellemare, Marc F. and Casey J. Wichman, "Elasticities and the Inverse Hyperbolic Sine Transformation," *Oxford Bulletin of Economics and Statistics*, 2020, *82* (1), 50–61.

Boschma, Ron, "Relatedness as driver of regional diversification: a research agenda," *Regional Studies*, mar 2017, *51* (3), 351–364.

Breschi, S. and F. Lissoni, "Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows," *Journal of Economic Geography*, 2009, *9* (4), 439–468.

Breschi, Stefano and Camilla Lenzi, "Co-invention networks and inventive productivity in US cities," *Journal of Urban Economics*, 2016, *92*, 66–75.

CAAC, "Civil Aviation Administration of China, Website."

Cao, Zhan, Ben Derudder, Liang Dai, and Zhenwei Peng, "'Buzz-and-pipeline' dynamics in Chinese science: the impact of interurban collaboration linkages on cities' innovation capacity," *Regional Studies*, 2021, pp. 1–17.

Carlino, Gerald and William Kerr, "Agglomeration and Innovation," *NBER Working Paper 20367*, 2014.

Correia, Sergio, "Singletons, Cluster-Robust Standard Errors and Fixed Effects: A Bad Mix *," 2015.

Cui, Jingbo, Tianqi Li, and Zhenxuan Wang, "High-Speed Railway and Collaborative Innovation: Evidence from University Patents in China," *SSRN Electronic Journal*, 2020.

D'Este, P., F. Guy, and S. Iammarino, "Shaping the formation of university-industry research collaborations: what type of proximity does really matter?," *Journal of Economic Geography*, 2013, *13* (4), 537–558.

Dijkstra, E. W., "A note on two problems in connexion with graphs," *Numerische Mathematik*, 1959, *1* (1), 269–271.

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

Dong, Xiaofang, "High-speed railway and urban sectoral employment in China," *Transportation Research Part A: Policy and Practice*, 2018, *116*, 603–621.

_ , Siqi Zheng, and Matthew E. Kahn, "The role of transportation speed in facilitating high skilled teamwork across cities," *Journal of Urban Economics*, 2020, *115*, 103212.

Duranton, Gilles, Peter M. Morrow, and Matthew A. Turner, "Roads and trade: Evidence from the US," *Review of Economic Studies*, 2013, *81* (2), 681–724.

Ellison, Glenn, Edward L. Glaeser, and William R. Kerr, "What causes industry agglomeration? Evidence from coagglomeration patterns," *American Economic Review*, 2010, *100* (3), 1195–1213.

Essletzbichler, Jürgen, "Relatedness, Industrial Branching and Technological Cohesion in US Metropolitan Areas," *Regional Studies*, 2015, *49* (5), 752–766.

Faber, Benjamin, "Trade integration, market size, and industrialization: Evidence from China's national trunk highway system," *Review of Economic Studies*, 2014, *81* (3), 1046–1070.

Fang, Hanming, Quanlin Gu, Wei Xiong, and Li-An Zhou, "Demystifying the Chinese Housing Boom," *NBER Working Paper 21112*, 2015.

Fleming, Lee, "Recombinant uncertainty in technological search," *Management Science*, 2001, *47* (1), 117–132.

Florida, Richard, Patrick Adler, and Charlotta Mellander, "The city as innovation machine," *Regional Studies*, 2017, *51* (1), 86–96.

Gao, Jian, Bogang Jun, Alex 'Sandy' Pentland, Tao Zhou, and César A. Hidalgo, "Spillovers across industries and regions in China's regional economic diversification," *Regional Studies*, 2021, pp. 1–16.

Gao, Yanyan and Jianghuai Zheng, "The impact of high-speed rail on innovation: An empirical test of the companion innovation hypothesis of transportation improvement with China's manufacturing firms," *World Development*, 2020, *127*, 104838.

_ , Shunfeng Song, Jun Sun, and Leizhen Zang, "Does High-Speed Rail Really Promote Economic Growth? Evidence from Chinaas Yangtze River Delta Region," *SSRN Electronic Journal*, 2018.

Goldewijk, Klein Kees, Arthur Beusen, and Peter Janssen, "Long-term dynamic modeling of global population and built-up area in a spatially explicit way: HYDE 3.1," *The Holocene*, 2010, *20* (4), 565–573.

_ , _, Gerard Van Drecht, and Martine De Vos, "The HYDE 3.1 spatially explicit databaseof human-induced global land-use change over the past 12,000 years," *Global Ecology and Biogeography*, 2011, *20* (1), 73–86.

Guo, Huanxiu, Cheng Chen, Xiaofang Dong, and Changmin Jiang, "The evolution of transport networks and the regional water environment: the case of Chinese high-speed rail," *Regional Studies*, 2021, pp. 1–27.

Guo, Qi and Canfei He, "Production space and regional industrial evolution in China," *GeoJournal*, 2017, *82* (2), 379–396.

Hausmann, Ricardo, Jason Hwang, and Dani Rodrik, "What you export matters," *Journal of Economic Growth*, 2007, *12* (1), 1–25.

He, Canfei, Shengjun Zhu, and Xin Yang, "What matters for regional industrial dynamics in a transitional economy?," *Area Development and Policy*, 2017, *2* (1), 71–90.

Henderson, Vernon, Adam Storeygard, and David N. Weil, "A bright idea for measuring economic growth," in "American Economic Review," Vol. 101 2011, pp. 194–199.

Heuermann, Daniel F and Johannes F Schmieder, "The effect of infrastructure on worker mobility: evidence from high-speed rail expansion in Germany," *Journal of Economic Geography*, 2019, *19* (2), 335–372.

Hidalgo, C. A., B. Winger, A. L. Barabási, and R. Hausmann, "The product space conditions the development of nations," *Science*, 2007, *317* (5837), 482–487.

Hidalgo, César A., Pierre Alexandre Balland, Ron Boschma, Mercedes Delgado, Maryann Feldman, Koen Frenken, Edward Glaeser, Canfei He, Dieter F. Kogler, Andrea Morrison, Frank Neffke, David Rigby, Scott Stern, Siqi Zheng, and Shengjun Zhu, "The Principle of Relatedness," in "Springer Proceedings in Complexity," Springer, 2018, pp. 451–457.

Hong, Wuyang and Mo Su, "Influence of Rapid Transit on Accessibility Pattern and Economic Linkage at Urban Agglomeration Scale in China," *Open Geosciences*, 2019, *11*, 804–814.

Hornbeck, Richard and Martin Rotemberg, "Railroads, Reallocation, and the Rise of American Manufacturing," *NBER Working Paper 26594*, 2019.

Hsiao, Cheng, H. Steve Ching, and Shui Ki Wan, "A panel data approach for program evaluation: Measuring the benefits of political and economic integration of Hong Kong with Mainland China," *Journal of Applied Econometrics*, 2012, *27* (5), 705–740.

Huang, Yue and Yebin Wang, "How does high-speed railway affect green innovation efficiency? A perspective of innovation factor mobility," *Journal of Cleaner Production*, 2020, *265*, 121623.

HYDE, "History Database of the Global Environment - the Netherlands Environmental Assessment Agency (PBL)," 2020.

Inoue, Hiroyasu and Kentaro Nakajima, "The Impact of the Opening of High-Speed Rail on Innovation," *RIETI Discussion Paper Series 17-E-034*, 2017.

Jaffe, A. B., M. Trajtenberg, and R. Henderson, "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations," *Quarterly Journal of Economics*, 1993, *108* (3), 577–598.

Jiao, Jingjuan, Jiaoe Wang, Fangni Zhang, Fengjun Jin, and Wei Liu, "Roles of accessibility, connectivity and spatial interdependence in realizing the economic impact of high-speed rail: Evidence from China," *Transport Policy*, 2020, *91*, 1–15.

Jin, Mengjie, Kun Chin Lin, Wenming Shi, Paul T.W. Lee, and Kevin X. Li, "Impacts of high-speed railways on economic growth and disparity in China," *Transportation Research Part A: Policy and Practice*, 2020, *138*, 158–171.

Ke, Xiao, Haiqiang Chen, Yongmiao Hong, and Cheng Hsiao, "Do China's high-speed-rail projects promote local economy?—New evidence from a panel data approach," *China Economic Review*, 2017, *44*, 203–226.

Komikado, Hiroshi, So Morikawa, Ayushman Bhatt, and Hironori Kato, "High-speed rail, inter-regional accessibility, and regional innovation: Evidence from Japan," *Technological Forecasting and Social Change*, 2021, *167*, 120697.

Kulkarni, Rajendra, Kingsley E. Haynes, R. R. Stough, and James D. Riggle, "Revisiting Night Lights as Proxy for Economic Growth: A Multi-Year Light Based Growth Indicator (LBGI) for China, India and the U.S.," *SSRN Electronic Journal*, 2012.

Lawrence, Martha, Richard Bullock, and Ziming Liu, "China's High-Speed Rail Development," *World Bank Publications*, 2019.

Li, Xuecao, Yuyu Zhou, Min Zhao, and Xia Zhao, "A harmonized global nighttime light dataset 1992–2018," *Scientific Data*, 2020, *7* (1).

Li, Yifan, "China High Speed Railways and Stations," *Harvard Dataverse, V1*, 2016.

Lin, Yatang, "Travel costs and urban specialization patterns: Evidence from China's high speed railway system," *Journal of Urban Economics*, 2017, *98*, 98–123.

_ , Yu Qin, and Zhuan Zie, "International Technology Transfer and Domestic Innovation: Evidence from the High-Speed Rail Sector in China," *CEP Discussion Papers*, 2015.

Lucas, Robert E., "Making a Miracle," *Econometrica*, 1993, *61* (2), 251.

Ma, Yong, Hongxia Su, Qian Jin, Wei Feng, Jianuo Liu, and Wenying Huang, *The General History of Chinese Tourism Culture*, SCPG PUBLISHING CORPORATION, 2016.

Marx, Matt, Deborah Strumsky, and Lee Fleming, "Mobility, skills, and the michigan noncompete experiment," *Management Science*, 2009, *55* (6), 875–889.

Melander, Eric, "Transportation Technology, Individual Mobility and Social Mobilisation," *CAGE Online Working Paper Series*, 2020, (471).

Mellander, Charlotta, José Lobo, Kevin Stolarick, and Zara Matheson, "Night-Time Light Data: A Good Proxy Measure for Economic Activity?," *PLOS ONE*, 2015, *10* (10), e0139779.

Miguelez, Ernest, "Collaborative patents and the mobility of knowledge workers," *Technovation*, 2019, *86-87*, 62–74.

Moretti, Enrico, "The Effect of High-Tech Clusters on the Productivity of Top Inventors," *NBER Working Paper 26270*, 2019.

Owen-Smith, Jason and Walter W. Powell, "Knowledge Networks as Channels and Conduits: The Effects of Spillovers in the Boston Biotechnology Community," *Organization Science*, 2004, *15* (1), 5–21.

Perlman, Elisabeth Ruth, "Dense Enough To Be Brilliant: Patents, Urbanization, and Transportation in Nineteenth Century America," *CEH Discussion Papers*, 2015, (036).

Petralia, Sergio, Pierre Alexandre Balland, and Andrea Morrison, "Climbing the ladder of technological development," *Research Policy*, 2017, *46* (5), 956–969.

Qin, Yu, "'No county left behind?' The distributional impact of high-speed rail upgrades in China," *Journal of Economic Geography*, 2017, *17* (3), 489–520.

Redding, Stephen J. and Matthew A. Turner, "Transportation Costs and the Spatial Organization of Economic Activity," in "Handbook of Regional and Urban Economics," Vol. 5, Elsevier B.V., 2015, pp. 1339–1398.

Rigby, David L., "Technological Relatedness and Knowledge Space: Entry and Exit of US Cities from Patent Classes," *Regional Studies*, 2015, *49* (11), 1922–1937.

Rodríguez-Pose, Andrés, "The revenge of the places that don't matter (and what to do about it)," *Cambridge Journal of Regions, Economy and Society*, 2018, *11* (1), 189–209.

Soete, Luc, "The impact of technological innovation on international trade patterns: The evidence reconsidered," *Research Policy*, 1987, *16* (2-4), 101–130.

Strambach, Simone and Benjamin Klement, "Cumulative and Combinatorial Micro-dynamics of Knowledge: The Role of Space and Place in Knowledge Integration," *European Planning Studies*, 2012, *20* (11), 1843–1866.

Tamura, Ryuichi, "The Effect of High-speed Railways on Knowledge Transfer: Evidence from Japanese Patent Citations," *Public Policy Review*, 2017, *13* (3), 325–342.

Teixeira, Aurora A. C., Paulo Santos, and Ana Oliveira Brochado, "International RD Co-operation between Low-tech SMEs: The Role of Cultural and Geographical Proximity," *European Planning Studies*, 2008, *16* (6), 785–810.

Tubiana, Matteo, Ernest Miguelez, and Rosina Moreno, "In knowledge we trust: learning-by-interacting and the productivity of inventors," *AQR Working Papers*, 2020, *2012005.*

Twitchett, Denis and Frederick W. Mote, *The Cambridge History of China 2: The Ming Dynasty, 1368 — 1644, Part 2*, Vol. 8, Cambridge University Press, 1998.

van Eck, Nees Jan and Ludo Waltman, "How to normalize cooccurrence data? An analysis of some well-known similarity measures," *Journal of the American Society for Information Science and Technology*, 2009, *60* (8), 1635–1651.

Veugelers, Reinhilde, "The challenge of China's rise as a science and technology powerhouse," *Policy Contributions*, 2017.

WIPO, "World Intellectual Property Indicators, 2012," Technical Report 2012.

_ , "World Intellectual Property Indicators 2019," Technical Report, WIPO 2019.

Wong, Jason C Y, "Blue-sky Thinking: Connectivity Impacts on Regional Economies and Innovation in the United States *," 2019.

WorldMap, "CH$_H$*SRail*$_2$011 − *WorldMap*,^{JJ} 2011.

Wuchty, Stefan, Benjamin F. Jones, and Brian Uzzi, "The increasing dominance of teams in production of knowledge," *Science*, 2007, *316* (5827), 1036–1039.

Yin, Deyun, Kazuyuki Motohashi, and Jianwei Dang, "Large-scale name disambiguation of Chinese patent inventors (1985–2016)," *Scientometrics*, 2020, *122* (2), 765–790.

Zheng, Siqi and Matthew E. Kahn, "China's bullet trains facilitate market integration and mitigate the cost of megacity growth," *Proceedings of the National Academy of Sciences of the United States of America*, 2013, *110* (14), E1248–E1253.

Zhu, Shengjun, Canfei He, and Qian Luo, "Good neighbors, bad neighbors: local knowledge spillovers, regional institutions and firm performance in China," *Small Business Economics*, mar 2019, *52* (3), 617–632.

Zou, Wei, Liangheng Chen, and Junke Xiong, "High-speed railway, market access and economic growth," *International Review of Economics and Finance*, 2019.

Appendix: Additional Tables

Table 7: HSR stations with speed more than 200km per hour and Innovation

Dep. var. =	Patents per capita		
	(1)	(2)	(3)
Panel A	**OLS**	**OLS**	**OLS**
IHS Distance to Stations	-0.1194***	-0.1215***	-0.1314***
	[0.0371]	[0.0365]	[0.0383]
Average Night Light	0.6214***	0.6592***	0.6392***
	[0.1832]	[0.1808]	[0.1886]
Flight Passengers	0.5593***	0.5934***	0.4058**
	[0.1696]	[0.1498]	[0.1831]
R-squared	0.83	0.77	0.75
Panel B	**IV**	**IV**	**IV**
IHS Distance to Stations	-0.2497***	-0.2179**	-0.2132**
	[0.0899]	[0.0874]	[0.0861]
Average Night Light	0.5510***	0.6098***	0.5997***
	[0.1858]	[0.1781]	[0.1867]
Flight Passengers	0.5327***	0.5622***	0.3806**
	[0.1720]	[0.1579]	[0.1907]
First-Stage F-stat	23.26	23.15	23.08
Sample Size	3141	3105	2988
City FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Sample	Full	Second and Third	Third

Notes: The dependent variable is the number of patents divided by popula- tion. Distance to high speed rail stations contains the straight line distance in meters from the most populous point of a city to the closest high speed rail-way station with a speed more than 200km per hour. Average night light data controls for the economic activity of a city. We control for the average crop- land area and the average land used for grazing. The variable distance to aHSR station has been transformed using the inverse hyperbolic sine transfor- mation. Panel A contains the OLS results based on equation 1 and panel B theIV estimates based on equation 3. Clustered standard errors at the city levelare reported in the brackets.

Table 8: Total Transportation cost and technological specialization model

Dep. var. =	Entry
	(1)
IHS Total Transportation Cost	-0.0131***
	[0.0030]
R-squared	0.31
Sample Size	1629435
City*IPC FE	Yes
City*Year FE	Yes
Year*IPC FE	Yes

Notes: The dependent variable is binary and switches to 1 if a city has a comparative advantage in a specific technological field. The main independent variable is the total computed transportation cost from any most populous city point to the most populous city points that are specialized in a given technological field. We include city*ipc fixed effects, city*year fixed effects and year*ipc fixed effects. OLS model with clustered standard errors at the city level are reported in the brackets.