Institut de Recerca en Economia Aplicada Regional i Pública — Document de Treball 2022/20, 61 pàg.
Research Institute of Applied Economics — *Working Paper 2022/20, 61 pag.*

Grup de Recerca Anàlisi Quantitativa Regional — Document de Treball 2022/10, 61 pàg.
Regional Quantitative Analysis Research Group — *Working Paper 2022/10, 61 pag.*

“Trains of Thought: High-Speed Rail and Innovation in China”

Georgios Tsiachtsiras, Deyun Yin, Ernest Miguelez and Rosina Moreno

UBIREA
Research Institute of Applied Economics

AQR
Regional Quantitative Analysis Research Group

UBIREA Institut de Recerca en Economia Aplicada Regional i Pública
UNIVERSITAT DE BARCELONA

WEBSITE: www.ub-irea.com • CONTACT: irea@ub.edu

AQR
Grup de Recerca Anàlisi Quantitativa Regional
Regional Quantitative Analysis Research Group

WEBSITE: www.ub.edu/aqr/ • CONTACT: aqr@ub.edu

Universitat de Barcelona
Av. Diagonal, 690 • 08034 Barcelona

Any opinions expressed here are those of the author(s) and not those of IREA. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

Abstract

This paper explores the effect of the High Speed Rail (HSR) network expansion on local innovation in China during the period 2008-2016. Using exogenous variation arising from a novel instrument - courier's stations during the Ming dynasty, we find solid evidence that the opening of a HSR station increases cities' innovation activity. We also explore the role of inter-city technology diffusion as being behind the surge of local innovation. To do it, we compute least-cost paths between city-pairs, over time, based on the opening and speed of each HSR line, and obtain that an increase in a city's connectivity to other cities specialized in a specific technological field, through the HSR network, increases the probability for the city to specialize in that same technological field. We interpret it as evidence of knowledge diffusion.

JEL Classification: R40, O18, O30, O33.

Keywords: High speed rail, Innovation, Technology Diffusion, Patents, Specialization.

Georgios Tsiachtsiras: University of Bristol and University of Bath, United Kingdom. Email: georgios.tsiachtsiras@bristol.ac.uk

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

Ernest Miguelez: Univ. Bordeaux, CNRS, BSE, UMR 6060, Avenue Léon Duguit, 33608 Pessac, France and AQR-IREA, University of Barcelona, Spain. Email: ernest.miguelez@u-bordeaux.fr

Rosina Moreno: AQR-IREA, University of Barcelona, Barcelona, Spain. Email: rmoreno@ub.edu

Acknowledgements

The paper was previously circulated as "Rails and innovation: Evidence from China". We would like to express our gratitude to Ron Boschma, Riccardo Crescenzi, Giovanni Dosi, Francesco Lissoni, Alex Coad, Antonio Andreoni, Carolina Castaldi, Gaétan de Rassenfosse, Bart Los, Jouke van Dijk, Ruben Gaetani, Martijn Smit, Yuan Tian, Nicola Cortinovis, Michele Pezzoni, Zhiling Wang, Valentina Di Iasio, 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 Develop ment" (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. Finally, we would like to thank Wu Zhao and Yuan Quan for being research assistants during the process of data collecting and cleaning. Rosina Moreno and Ernest Miguelez acknowledge financial support provided by the Ministerio de Ciencia e Innovación for the project entitled 'Innovation, Environment and Social Justice (JUST GREEN-INN), PID2020-119096RB-I00. Georgios Tsiachtsiras acknowledges financial support from Agaur (FI-DGR). Any remaining errors are ours.

1 Introduction

Over the last decades, China has become a science and technology powerhouse, and ambitious to become a global innovation and technological leader by mid-21st century (The Economist, 2018, 2019; Veugelers, 2017). Since 2011, the China National Intellectual Property Administration (CNIPA) has systemathically outnumbered other patent offices, such as the Japan Patent Office (JPO) or the United States Patent and Trademark Office (USPTO), in terms of annual patent applications received (WIPO, 2021). If one concentrates on the international patents produced within China's territory, the country has recently overcome the Republic of Korea, reaching 14% of the world patent production, not far from other big patent producers (US, 21.1%; Japan, 21%; Europe, 23.9%) (WIPO, 2019).[1]

Yet, as in many other countries, the internal spatial distribution of innovation is highly skewed (Feldman and Kogler, 2010). For instance, for the 2011-2015 period, Beijing, Shanghai, and Shenzhen concentrated 52.2% of all international patents (WIPO, 2019). Concentration in space is a typical economic phenomenon, particularly among high-tech industries. It arises from the existence of agglomeration economies, such as larger market access, labour market pooling, or localized knowledge spillovers, among others things (Carlino and Kerr, 2014; Jaffe et al., 1993; Lucas, 1993; Moretti, 2021). Transportation infrastructure endowments facilitate the flow of goods, services, labor, and human capital, thus affecting the quality and quantity of human interactions (Agrawal et al., 2017), in turn impacting agglomeration economies and knowledge spillovers. This is potentially the case of the rollout of the High-Speed Rail (HSR) in China, which was first introduced in 2007 and rapidly developed during the following 15 years.

Against this backdrop, we put forward the question of whether the relation between HSR infrastructure investments affect the way people and firms interact, and may affect the geography of innovation. The HSR network allows longer commuting, hence changing market size and market opportunities for local innovators, and favoring better employer-employee matching (Moretti, 2021). Moreover, scientists and inventors see increased opportunities to collaborate and build new and larger teams, generating more novel ideas (Dong et al., 2020; Wuchty et al., 2007). Finally, it allows connecting locations far apart (Cao et al., 2021), in turn favouring the recombination of knowledge pieces across the space (Owen-Smith and Powell, 2004).

Our analysis finds solid evidence that the opening of a HSR station is associated with an

[1] International patents refer to patent families produced by organizations and inventors residing in mainland China but seeking to protect their inventions beyond their home office (the CNIPA). For more details, see WIPO (2019).

increase in the number of patents per capita of a city, for the period 2008-2016. We further argue that this relationship exists not only in large urban centers but also among second and third-tier cities, despite being more peripheral in the innovation system of Chinese cities.

Then, we explore one of the main mechanisms behind the surge in local patenting, and investigate to what extent the HSR roll-out affects knowledge diffusion across the space. Borrowing the idea of technological diversification from the branching literature (Boschma, 2017; Essletzbichler, 2015; Hidalgo et al., 2007; Hausmann et al., 2007; Rigby, 2015), we test whether the reduction in transportation cost, due to an infrastructure improvement, affects knowledge diversification towards domains in which the inter-connected cities are specialized in. We demonstrate that the probability of a city to branch into a new technological field is related to the current portfolio of the cities to which it connects to through the HSR network, even after accounting for the local stock of capabilities.

Our findings contribute to the literature in several ways. First, we provide fresh evidence on the relationship between HSR and local innovation in China, and propose a novel causal interpretation, which relies on historical couriers' stations during the Ming dynasty (1403–1644), in order to create exogenous variation for our main explanatory variable.[2] Further, we also run a difference-in-differences regression with staggered treatment in which we can control for pre-trends prior to the arrival of the HSR network, and a falsification test between our instrument and innovation performance before the creation of the HSR network.

Second, we investigate whether the way in which HSR affects local innovation is through city-to-city knowledge diffusion. Yet, we depart from traditional measures of diffusion, such as patent citations, as they are not fully reliable in our context. Again, in this paper we investigate the role of the reduction of transportation costs, due to the expansion of the HSR network, on 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 to through HSR.

Finally, we investigate the above-mentioned issues through the analysis of an under-exploited dataset of fully geocoded and disambiguated Chinese inventors (Yin et al., 2020), which covers the entire universe of patent applications in CNIPA (2020 edition).[3] Further,

[2] Previous studies on the role of transport infrastructure use either an ancient transportation network (dated 1934 or 1961-1962) as instrument (Baum-Snow et al., 2017; Dong et al., 2020; Hanley et al., 2021; Zheng and Kahn, 2013), or a difference-in-difference approach (Dong, 2018; Gao et al., 2018; Gao and Zheng, 2020; Lin, 2017; Qin, 2017), among others techniques (Banerjee et al., 2020; Dong, 2018; Faber, 2014; Gao et al., 2018; Hsiao et al., 2012; Ke et al., 2017).

[3] Many papers in the literature so far, only use certain parts of the CNIPA database, like Cui et al. (2020),

in order to create consistent control variables for the same sample of Chinese cities over time, we complement our dataset with global raster files of population, cropland, grazing area (Goldewijk et al., 2017), as well as night light data (Li et al., 2020).

The rest of the paper is organized as follows: Section 2 reviews the relevant literature. Section 3 presents the empirical strategy and section 4 describes the data sources and variables construction. Section 5 presents the main results. Last section (6) concludes.

2 Literature Review

The paper speaks to different strands of literature. First, it contributes to the scholarly work linking transport infrastructures to economic performance (Aschauer, 1989; Fogel, 1962; Garcia-Mila and McGuire, 1992; Munnell, 1992). A causal interpretation of this relationship has been the workhorse of much recent research. As a mode of example, Duranton and Turner (2012) study employment growth in US cities as a function of the city's stock of highways (1983-2003). Exogenous variation comes from relying on three different instruments based on historical information on US railroads and the interstate highway system. For the case of China, Baum-Snow et al. (2020) investigate the impact of highways on economic activity and population, around the year 2010. Identification in this case comes from the variation of the 1962 road network, which pre-dates China's transformation of recent decades. This same instrument is used to identify the effect of Chinese highway and railroad networks on the relative decentralization of population and industrial production from central cities to suburban and ex-urban areas (Baum-Snow et al., 2017). Also for the case of China, Banerjee et al. (2020) deal with the problem of the endogenous location of transportation networks by taking advantage of the fact that they tend to connect historical cities, and conclude that being close to the network had a positive level effect on GDP per capita, although this is not large.

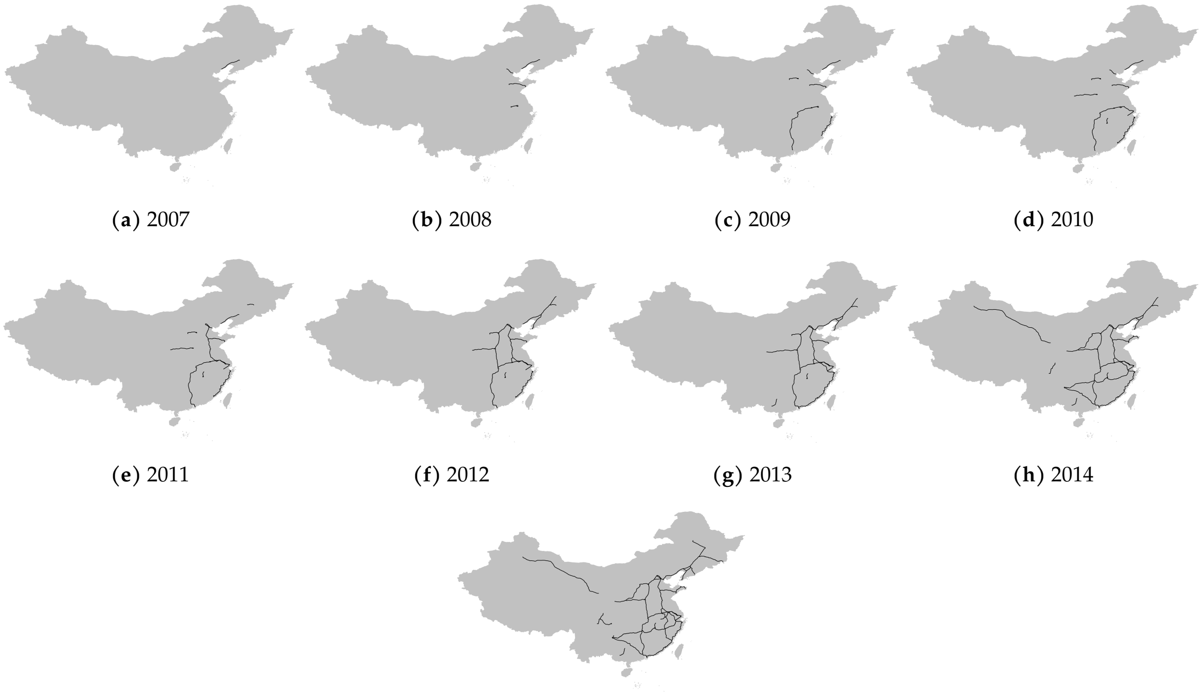
The development of the railroad networks have historically attracted the attention of scholars, as, since their inception, they became the dominant form of freight transport, and cities around railroad lines prospered economically (Donaldson and Hornbeck, 2016; Fogel, 1962). More recently, the roll-out of the HSR and its relationship with several economic outcomes has risen similar interest, establishing a parallelism with historical studies on the roll-out of the conventional railroad network. This is due to the combination of rapid expansion (in certain countries) plus larger and faster transport capacity, in this case of people - with affections in commuting and human interaction costs. For instance, Heuer-

who rely only on university patents.

mann and Schmieder (2019) explore the expansion of the 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. Ahlfeldt and Feddersen (2018) analyze the economic impact of the German HSR connecting Cologne and Frankfurt on the GDP of the three counties with intermediate stops, finding a causal positive effect of the HSR on GDP, although the strength of the effect declines with travel time between counties.

The roll-out of the HSR in China is of particular interest, because of the speed of its implementation: the first HSR line was introduced in the country in 2007, and the rest of the network developed rapidly over the following 15 years, becoming the longest in the world. It is designed for speeds of 250–350 km/h (155–217 mph), and it aims to complement existing transportation networks (Lin, 2017). Currently, the HSR network moves around 1.7 billion passengers per year (Lawrence et al., 2019). Figure 1 presents the expansion of the HSR network from 2007 to 2015.

Figure 1: Expansion of the HSR network in China



(a) 2007

(b) 2008

(c) 2009

(d) 2010

(e) 2011

(f) 2012

(g) 2013

(h) 2014

(i) 2015

Notes: Only lines with an average speed of 250 kilometers per hour and more are included. Source: authors elaboration based on Li (2016) and Wendy et al. (2012).

The opening of the HSR in China has been linked to regional economic growth (Gao et al., 2018; Yu, 2021; Zou et al., 2019). The way HSR impacts regional outcomes is through facilitating the movement of people and thus changing the size of markets and market access (Gao et al., 2018; Lin, 2017; Qin, 2017). HSR has also important distributional effects between connected and not connected areas (Gao et al., 2018; Qin, 2017). The effect of HSR is far from being homogeneous, though. HSR's impact may depend on the location, route, and region of the cities that it connects (Ke et al., 2017). In general, more industrialized cities, and those capable to absorb enough labor and stocked with better infrastructure are more likely to gain from better HSR connections (Ke et al., 2017). Yet, secondary cities can also gain when connected to star cities through the HSR network (Dong et al., 2020). Not only territorial, but also industry differences may exist. Dong (2018) finds that the HSR network has a different effect on employment growth among different industries, with significant impact on employment of the retail/wholesale and hotel/food industries, while no relevant effects in other sectors. Meanwhile, Duan et al. (2021), looking at venture capital investment data in China, find that small cities, high-tech industries, and younger firms are the ones mostly benefited by the implementation of the HSR, which seems to give an important role to information transmission, and not only market access.

Indeed, most of the mentioned literature so far focuses on market access as the mechanism behind the relation between transportation infrastructure and economic outcomes, leaving a minor role for knowledge diffusion (Agrawal et al., 2017; Bottasso et al., 2022; Tsiachtsiras, 2020). But trains not only transport people and goods, but also ideas that can be learned elsewhere or even be recombined with local ones. From a historical perspective, Andersson et al. (2021) identify the causal contribution of the establishment of a railway network connection in 19th century Sweden, finding that it increased both the spatial spread of innovative activity and the pace of local innovation. Among the mechanisms, the authors not only consider that railroads induce innovation through local economic growth, but also by increasing the demand for innovation and the supply of innovation through knowledge spillovers, as well as by integrating local economies with external markets and facilitating the exchange of ideas between places. Coming to more recent times, Tamura (2017), Inoue and Nakajima (2017) and Komikado et al. (2021) investigate the effect of the HSR on innovation activities in Japan. They find that the opening of the Hokuriku Shinkansen line in 1997 facilitated knowledge diffusion, as measured by patent citations (Inoue and Nakajima, 2017; Tamura, 2017) and patent collaborations (Inoue and Nakajima, 2017). Furthermore, the existence of HSR stations has a positive and statistically significant association with knowledge productivity, measured as patent applications per employee in each Japanese prefecture (Komikado et al., 2021).

Moving to the Chinese case, Lin et al. (2015) find that railway firms gain in terms of knowledge spillovers from the foreign technology transfer, with nearby firms benefiting not only in terms of more patents, but also with a higher productivity and revenue growth. Yang et al. (2021) show that the cities without a HSR station benefit from the connection to the HSR in neighboring cities as they also experience an increase in their innovation activity, pointing to the effect of knowledge spillovers related to the HSR expansion not only in connected cities, but also to nearby ones. Likewise, it is shown that the HSR connection contributes to the increase of innovation activities of firms (Gao and Zheng, 2020), the scientific productivity and scientific knowledge diffusion of secondary cities (Dong et al., 2020), as well as the capacity to produce green innovations, thanks to the increase in the mobility of innovative factors (Huang and Wang, 2020).

Recently, a stream of research uses gravity models and finds that travel time reduction, thanks to the expansion of the HSR network, enhances patent collaborations between inventors (Sun et al., 2021; Yao and Li, 2022), between scientists in star metropolis with second tier cities (Dong et al., 2020), between firms and universities (Cui et al., 2020), and between enterprises in different locations (Hanley et al., 2021), and in turn facilitates the diffusion of knowledge, as measured with patent citations across cities (Hanley et al., 2021). Following this strand of literature, we build an improved measure of transportation cost based on the least-cost path to explain knowledge diffusion across cities, also using co-patenting and cross-city citation data as dependent variables. Yet, due to the multiple methodological issues of these variables (Jaffe and de Rassenfosse, 2017), we relegate this analysis to the online appendix, and address the HSR-diffusion relationship as follows.

In specific, we go beyond the use of patent collaborations and citations as a measure of knowledge flows, and build on the branching literature on technological diversification (Boschma, 2017; Essletzbichler, 2015; Hidalgo et al., 2007; Hausmann et al., 2007; Rigby, 2015), which has investigated to what extent a city diversifies in a new technology (develops comparative advantage) as a function of the presence of related technologies in that city, in line with the principle of relatedness (Gao et al., 2021; Guo and He, 2017; He et al., 2017; Hidalgo et al., 2018; Zhu et al., 2019). For instance, Balland et al. (2019) provide evidence that relatedness and knowledge complexity affect the probability of a region to specialize in a given technological field, across European regions. Petralia et al. (2017), at the country level, investigate how the proximity of countries' existing capabilities to a certain technology is associated with the probability to specialize in such technology. None of these studies, however, address the issue of diversification thanks to the connections to the outside world, which is one focus of our analysis. Yet, this is likely to be the case, at the level of cities, regions, or countries. As a mode of example, Bahar et 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 inventors originating in their countries of origin. Bahar et al. (2014) establish that neighboring countries are very similar in their patterns of comparative advantage, a similarity that decays with distance, suggesting knowledge diffusion as a potential underlying mechanism behind their main findings, although without testing it empirically. Economic geography papers have also looked into diversification induced by external actors, such as, respectively, high-skilled immigrants, non-local entrepreneurs, and co-inventors (Miguelez and Morrison, 2022; Neffke et al., 2017; Whittle et al., 2020; Elekes et al., 2019). Finally, our paper also relates to Gao et al. (2021), who explore spillovers across industries and regions in China's regional economic diversification at the province level. In their analysis, they use the HSR network as an instrument in a differences-in-differences design to explore whether the timing of the roll-out of the HSR is associated with industrial convergence between provinces. Yet, the authors do not explicitly account for the way in which the HSR affects cross-city knowledge diffusion. In this paper, we borrow from these ideas and study to what extent the probability of a city to specialize in a new technological field is related to the specialization patterns of the cities to which the city connects to through the HSR.

3 Empirical Strategy

3.1 Baseline approach

To begin with, we estimate the effect of the HSR on the innovation performance of Chinese cities as follows:

$$Y_{it} = \alpha + \beta H_{it-1} + \gamma_i + \delta_t + \zeta X_{it-1} + \epsilon_{it} \tag{1}$$

where Y_{it} is the number of patents per capita in city i and time period t. The main variable of interest, H_{it-1}, built as the straight line distance of any city point with the highest population value to the closest HSR station, as computed in the previous period.[4] We include city fixed effects, γ_i, and year fixed effects, δ_t, as well as several control variables, X_{it-1}, also one-year lagged. First, in order to control for economic activity at the city level, we rely on night light data (Henderson et al., 2011; Mellander et al., 2015). Kulkarni et al. (2012) report evidence that for only a very low count of prefectures (between 5 to 10%), the night light data is not a good proxy for determining local GDP in the case of China.

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

We further control for the cropland and grazing density within the city.[5] The purpose of these two controls is to capture local agglomeration economies and urbanization features of the areas, which may correlate with innovation (Beckers et al., 2020). Also, we include the number of flight passengers, in line with the literature (Dong et al., 2020), to take into consideration the airplane as an alternative transportation mean. Finally, we control for the fact that a city may belong to a high technological zone, since companies located in these areas not only benefit from better infrastructure and access to talent but also can receive special incentives, such as lower taxation (Wang and Feng, 2021; Zhuang and Ye, 2020). We estimate our model by OLS, clustering the standard errors at the city level, and log-transforming the main independent variable.

Yet, our baseline OLS estimates could well be compromised due to endogeneity issues. This comes from the fact that the placement of the HSR stations is not randomly established, and consequently, transportation investments may be correlated with the outcomes of interest (Andersson et al., 2021) or assigned through the existing political process (Redding and Turner, 2015).[6] For these reasons, we adopt an identification strategy based on instrumental variables (IV), as explained next.

3.2 Identification Strategy: Historical Couriers' Routes and Stations

Existing Literature

In order to deal with the endogenous nature of the network infrastructure, common approaches in the literature are the computation of least-cost paths between cities or the use of an old infrastructure network as instruments for railway access (Büchel and Kyburz, 2020; Andersson et al., 2021). The least-cost path instrument allows the researchers to confront endogeneity concerns related to the placement of railway lines exploiting quasi-random variation coming from geographic features such as the slopes, the river crossings and the land cover as inputs for the computation of the least-cost paths. However, the selection of the geographical elements and their importance is arbitrary and endogenous to the identification strategy, that is, the econometrician takes a number of decisions on the importance of each element, which in turn are likely to be correlated with the placement of the least-cost path and the economic outcomes.

[5] These two variables are computed as follows: we aggregate, by city, the value of each variable for every pixel using the HYDE raster files (Goldewijk et al., 2017). Then, we divide the sum by the area of the city, in km^2.

[6] The territorial targeting of public resources for strategic electoral reasons has always been at the centre of significant scholarly work (Golden and Min, 2013).

Using an old infrastructure network presents its own drawbacks (Baum-Snow et al., 2017). These are related to the fact that an old infrastructure network may be correlated with unobserved variables that influence the recent infrastructure network and the evolution of the cities. As a partial solution, researchers include controls on the evolution of the city economies to strengthen their argument that the old transport network components predict recent outcomes solely through their influence on the location and configuration of the modern transportation network (Baum-Snow et al., 2017).

Couriers' Routes

We follow a different approach and rely on couriers' routes and their stations during the Ming dynasty (1368-1644) to create an exogenous variation for our analysis. We argue that the routes represent the least-cost paths used by the couriers to deliver their messages among Chinese cities. The advantage of this approach is that we do not have to make any assumption on the effect of topography on the least-cost paths, as we do not use geographical characteristics or the location of the big urban centers to draw least-cost paths. In our case, instead of using an old infrastructure network, we rely only on the least-cost routes identified and used by the couriers, as they represent the fastest ways for travelling across cities. Moreover, the time span of our instrument clearly pre-dates the important changes that China experienced during the 20th century.

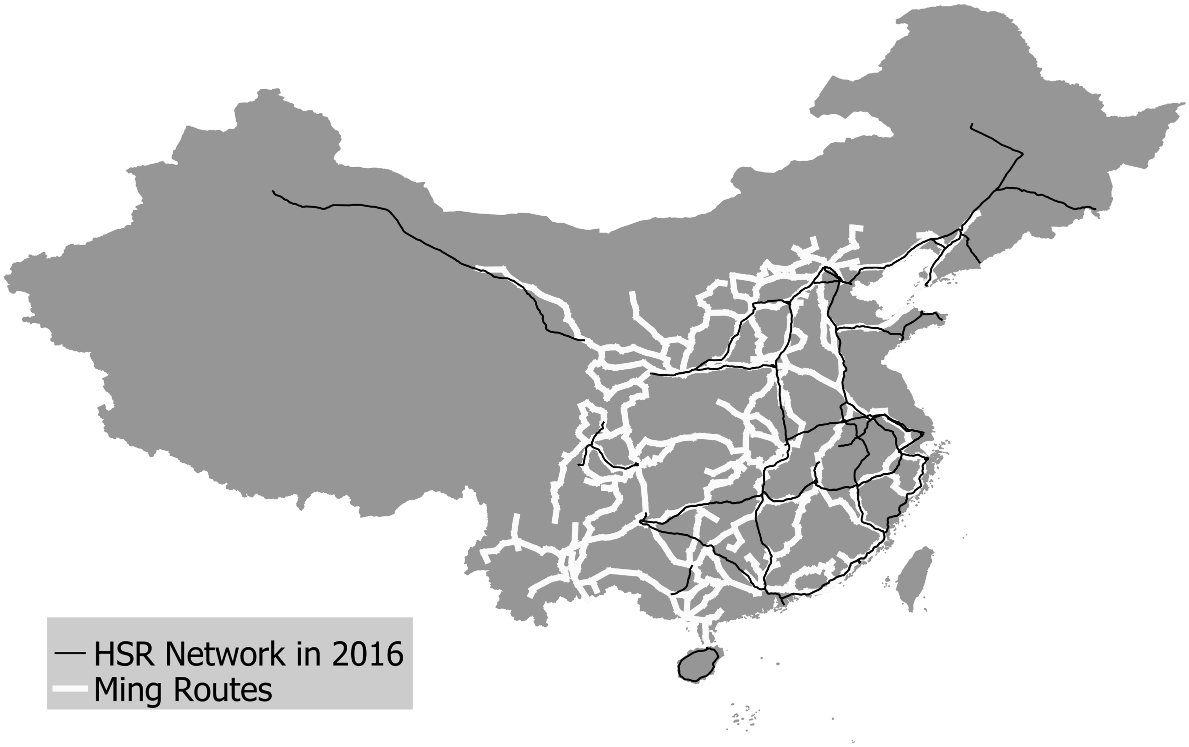
Twitchett and Mote (1998) describe in detail the organization of the courier services and provide illustrative examples. There were 1936 operational stations established in mainland China, covering the major routes of the Ming dynasty. They argue that it is reasonable to assume that these routes were the fastest ones since there was a penalty for the couriers exceeding the time limits to deliver the mails. For exceeding the time limit by a day, a courier was liable to a beating of twenty strokes, plus an additional stroke for every three days beyond that, to a maximum of sixty.[7] Zhang et al. (2021) document that when a courier reached a station, the arrival time had to be written down on the notebook and there was a punishment if the courier was late. More details about the success of the Ming communication network are included in the stories of foreign visitors whose entire presence in China was overseen by a series of well-connected courier stations spaced apart by

[7] Twitchett and Mote (1998) provide stories which could serve as anecdotal evidence: "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 gave 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" (pp. 586-588). Using the couriers' routes they arrived in Peking two days before their travel permit expired.

an average of 30 kilometers (Brook, 1998).[8]

The Ming dynasty is possibly the most appropriate time period to build our instrument, because the distribution system of couriers was renovated and reached at its peak in terms of efficiency. In the years after the Ming dynasty, and especially during the Qing dynasty, the courier system started to decline because of the wars (Ma et al., 2016). Figure 2 shows the courier routes during Ming dynasty, alongside the lines of the HSR network in 2016.

Figure 2: HSR lines in 2016 and Ming routes



Notes: Own construction using data from Li (2016) and Wendy et al. (2012).

We compute the straight line distance from any city point with the highest population value to the closest courier station. Since the instrument is static, we multiple it with time dummies (Andersson et al., 2021; Melander, 2020). We expect a positive relation between the instrument and the straight line distance of the most populated point in a city to the closest HSR station. A city which is far away from the courier stations during Ming dynasty

[8] In appendix C, Tables C.7 and C.8, respectively, as a robustness test we use the distance to couriers' routes instead of the distance to couriers' stations, for our IV. This robustness check should mitigate concerns related to the location of the stations along the routes.

should also be far away from the HSR stations in more recent periods.

Our first stage equation is as follows:

$$H_{it-1} = \alpha + \sum_{y} \kappa_t^y (Distance\ to\ Courier\ Station)_i + \gamma_i + \delta_t + \zeta X_{it-1} + \epsilon_{it} \tag{2}$$

where κ_t^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. We then end up with eight instruments, one for each year after the introduction of the HSR network (Andersson et al., 2021). Distance to courier stations is transformed using logs. In a second stage, our main equation is:

$$Y_{it} = \alpha + \beta \hat{H}_{it-1} + \gamma_i + \delta_t + \zeta X_{it-1} + \epsilon_{it} \tag{3}$$

where $\hat{H}_{it-1}$ 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 equation 1.

Testing for the exclusion restriction

It is true that an old infrastructure network could correlate with recent outcomes of interest (Baum-Snow et al., 2017). The routes of the couriers do not reflect an old infrastructure network but rather the fastest routes across cities. Nevertheless, we provide additional empirical evidence to check for the validity of our instrument.

We begin by applying, as a robustness exercise, a falsification test in line with Autor et al. (2013) and Dix-Carneiro et al. (2018) to make sure that our estimating results are not affected by pre-trends (reverse causality) the years immediately prior to the expansion of the HSR network. For this reason, we explore the impact of the future HSR network on past innovation activity.

Furthermore, while the instrument highly correlates with our focal variable in the first stage, we also need to ensure that it meets the exclusion restriction, that is, that our instrument does not affect the outcome directly, or through other variables. In this sense, we argue that our list of controls (population, night light, urbanization,...) blocks off the causal channel from courier's stations during the Ming dynasty. In our robustness checks, we also include two additional control variables related to the Ming dynasty that account for economic activity in those times, in order to make sure the instrument meets the exclusion restriction: (1) the share of people that took the entry exams in each Chinese city (Wendy et al., 2012) and (2) the average population growth of cities during the Ming dynasty (Goldewijk et al., 2017).

3.3 HSR and the diffusion of knowledge across cities

As surveyed above, knowledge diffusion and idea recombination is in large part behind the surge of local patenting. Consequently, in this paper we investigate the extent to which the HSR roll-out affects knowledge diffusion across the cities in China. We use cross-city gravitational models and test how the reduction of travelling cost between pairs of cities influences city-pair counts of patent citations (Jaffe et al., 1993) and collaborations (Hanley et al., 2021). Different from previous studies, and as explained in section 4.3, we compute the yearly cost of travelling from one city to another via the HSR, rail and road network as the least-cost path of going from a given city point with the highest population value to all the other cities' most populated points, applying the Dijkstra (1959) algorithm. Unfortunately, the use of patent citations and patent collaborations does not come without important drawbacks. First, recent studies have started to criticize the use of citations as a measure of knowledge flows as flawed (Arora et al., 2018; Jaffe and de Rassenfosse, 2017). On the other hand, patent collaborations from CNIPA come with its own problems, too. Chiefly, instead of inventors' addresses, CNIPA only registers the address of the first applicant for each patent (Yin et al., 2020), thus making impossible to compute measures of co-inventorship across cities. We partially solve the former issue by attributing the addresses of non-first applicants collecting their address information when they appear as first applicants in other patents. We measure co-patenting this way, but not co-invention. For the latter, we rely on USPTO patent equivalents when inventors reside in mainland China, getting therefore information on inventors' addresses. Unfortunately, USPTO equivalents are only a small proportion of all patents in China.

For all these reasons, we prefer to relegate the gravity analysis to the online appendix D, and suggest here a different approach to study city-to-city knowledge diffusion generated by the HSR. Particularly, we explore the effect of the HSR network on the technological diversification of cities. We investigate whether the reduction of transportation costs, due to the expansion of the HSR network, has an impact on the technological specialization of cities, by exposing these cities to the knowledge produced in the areas to which it connects to. With this objective in mind, we estimate the following equation:

$$Entry_{ict} = \alpha + \beta EEK_{ict-1} + \rho RelDen_{ict-1} + \lambda_{it} + \phi_{ct} + \omega_{ic} + \epsilon_{ict} \tag{4}$$

where $Entry_{ict}$ is a binary variable switching to 1 if the city i specializes in a given technology c in year t, conditional on not being specialized in that technology the previous year, $t-1$. Specialization is measured using the relative technological advantage (RTA) index for each city and technology:

$$RTA_{ict} = \frac{pat_{ict}/\sum_c pat_{it}}{\sum_i pat_{ct}/\sum_c \sum_i pat_t} \tag{5}$$

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). If the city preserves a RTA index larger or equal than 1 in the following years, $Entry_{ict}$ is populated with missing values, since a city cannot enter again, by definition, before exiting. If a city has a RTA index lower than 1 in period t as well as in period $t-1$, it preserves the zero value.

We define our focal explanatory variable, *Exposure to External Knowledge*, EEK_{ict-1}, as follows:

$$EEK_{ict-1} = \sum_{j \neq i} \left(\frac{Patents_{ijct-1}}{Cost_{ijt-1}} * D\_RTA_{jct-1}\right) \tag{6}$$

where, for every year t, we compute:

- For each city i and technology c, we compute the number of patents in all the cities j in technology c, $Patents_{ijct-1}$
- We weight all patents j in technology c by the least-cost path of going from city i to city j (we divide patents of j by the cost of travelling from i to j), $\frac{Patents_{jt-1}}{Cost_{ijt-1}}$
- We multiply the resulting ratio of the two elements above by a dummy variable, D_RTA_{jct-1}, indicating if the city j has RTA in technology c in the period $t-1$, that is, if $RTA_{jct-1} > 1$
- We sum all the resulting computations for city i and technology c, which connect to all the other cities j's

We expect that a high value of this variable implies a high exposure to external knowledge thanks to the HSR. In other words, we claim that a better connectivity with the patent portfolio of the cities already specialised in technology c, will impact the likelihood of city i to specialize in that same technology c.[9]

[9] Results are also provided for a version of the EEK variable computed with fixed values of patents produced

Following the literature on technological diversification, we control for the degree of relatedness between any technology c and the technological portfolio of a city i, with a relatedness density indicator, $RelDen_{ict}$.[10].

Finally, our regression also includes the number of patents for every city and every technological field, the RTA of the city in technology c, city-year fixed effects, λ_{it}, and technology-year fixed effects, ω_{ic}. We cluster the standard errors at the city level. Following the literature, all our variables have been time-lagged one year.

4 Data sources and variable construction

4.1 Patent Data

Our main source of data is the Patent Office of the People's Republic of China (CNIPA), founded in 1980 (2020 edition). The assignment of patents to cities is based on 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 patents to cities based on the latitude and longitude of this address. Our maps of Chinese cities (in shapefile format) 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. Previous studies consider Chinese patent applications as of relatively low quality (Eberhardt et al., 2017). To deal with this, we also collect data on claims and forward citations, in order to account for innovation quality and impact in a set of robustness analyses.[11] We extract the data on claims from Google patents and citations from the EPO Worldwide Patent Statistical Database (PATSTAT).Thus, we weight our city level measure of innovation (number of patents per 10000 people) with quality indexes of such innovation that have been previously used in the literature: i) the number of patents weighted by their number of claims; ii) the number of patents weighted by their number of forward citations received within a 5-year window (Aghion et al., 2019); and iii) the number of patents that belong to the top-50 percent of the distribution sorted by forward citations, for every year and for every

in cities j in technology c, as well as of the D_RTA_{jc}. In particular, we use the patents of the period 1998-2007. In this way, variation comes only from the reduction in transportation costs from city i to city j. The results are provided in the Table 4 columns 4 to 6.

[10] Technical details on the construction of relatedness density indicator can be found in the Appendix A

[11] We could not recover citation data in 6,471 cases out of 4,105,238. However, we believe that it is unlikely that these missed cases affect our analysis as these are patents for military purposes. Regarding patent claims, we were not able to extract data in 239,353 patent applications out of 4,105,238. Again, we believe that this is unlikely to affect our results to a large extent.

technological class, within a 5-year time window (Tubiana et al., 2022).

Technological classes, or technologies, are defined using the first four digits of the IPC codes listed in patent applications. We restrict our analysis to technological classes that appear throughout all the years. Thus, our final database includes 600 technological classes.

4.2 Rail Data

We use three different sources to construct the HSR network. We extract data on HSR stations and lines of China's HSR System from the shapefiles in Li (2016) and Wendy et al. (2012), where we can find the most precise and accurate placement of lines and stations. We also extract data on the speed of each line in two different time periods (2011 and 2016). Finally, we obtain the opening years of the HSR stations from the official website of the National Railway Administration of the People's Republic of China.[12]

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 the straight line distance, in meters, from the city's most populous point to the closest HSR station.[13] To identify the most populous points we use the raster file of population in 2007 from HYDE database (Goldewijk et al., 2017). We transform the pixels into points for every city and we choose the pixel with the highest population value. An alternative approach would have been to choose the centroids of the polygons instead of the most populous points, but our choice makes the empirical setting more realistic, as the most populous points represent the center of the city to a better extent.

4.3 Transportation cost

To compute the accessibility of each city due to the expansion of the HSR network, we go through the following steps. First, we divide China into squared grids, each one having a side of 10 kilometers. We assume that the crossing of a grid requires a distance of 10 kilometers (equal to one side) to be covered.[14] Our source of variation over the years is the average travel speed of each HSR line. We rely on the speed of the lines to allocate costs to our grids. We have 3 categories of HSR lines, according to their travel speed: 200,

[12] www.12306.cn, last accessed May 2020.

[13] We have information about the latitude and the longitude of every HSR station.

[14] We make this assumption in order to express our transportation cost in time units. It is true that a passenger can stop in the middle of a grid without crossing it or to use the diagonal which is even longer than the side. For this reason, we make the assumption of 10 kilometers which is equal to one side of the grid.

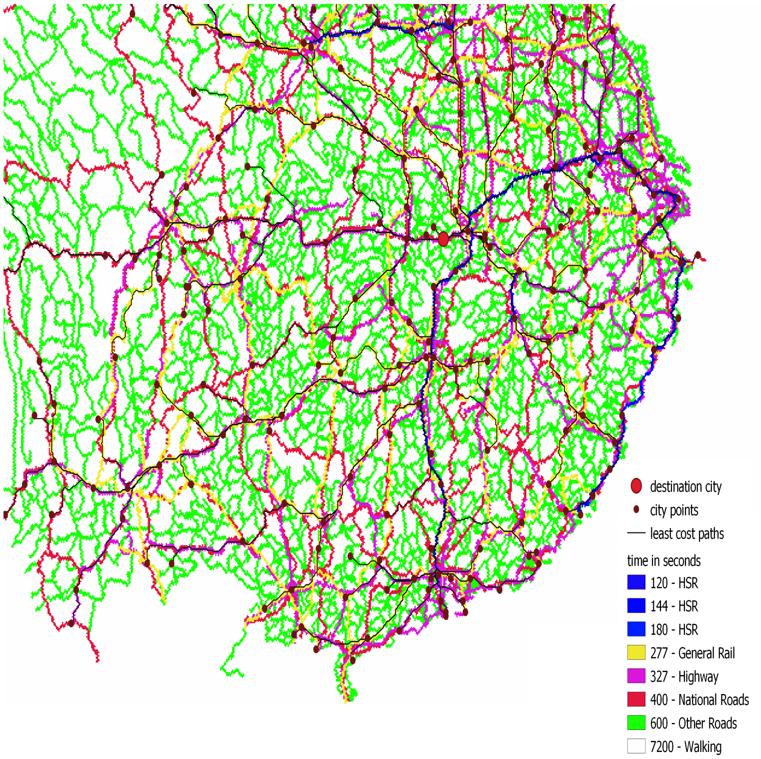
250 or 300 kilometers per hour, on average. Furthermore, we make use of shapefiles of the road network in 2009 and general rail network in 2005, extracted from Wendy et al. (2012). In total, we have 8 different transportation modes (HSR (300, 250, 200), General Rail, Highway, National Roads, Other Roads and Walking. Regarding the walking speed we make use of the value 5km/hour like in Mimeur et al. (2018).

Then, we compute the cost values using the following well know formula of time:

$$t = [d/s] * 3600 \tag{7}$$

where t is the time period, d is distance (always set to 10km) and s is the average travel speed for all the lines: HSR, general rail, roads and walking. We multiple it with 3600 to have the outcome in seconds. Technical details regarding the transportation costs can be found in the appendix A.

Figure 3: Example of least-cost paths



Notes: Souce: Authors' computations.

Later, we create our cost raster files, one for each year, which identify the cost of traveling through each cell (Büchel and Kyburz, 2020). Then, we apply the Dijkstra (1959) algorithm to compute the least-cost paths for a given city point with the highest population value to all the other cities' most populated points. This is a typical minimisation problem based on the least-cost surface, which selects the optimal route.

Figure 3 presents an example of the least-cost paths obtained for the transportation network in 2010. In this example, the focal city i is in the center with a deep red colour defined as a destination city. Then, the algorithm finds the least-cost path from every city's j most populated point, black smaller dots, to the focal city i (to its most populated point). Blue lines are the HSR lines, yellow lines are the general rail network, pink lines are the highways, red lines are the national roads, green lines all the other smaller roads, and white areas can only be crossed via walking. Black lines are the least-cost paths along the transportation networks. The result is a vector of 123,201 least-cost paths, for every year, which can be summarized in the following cost matrix:

$$C_t = \begin{vmatrix} Cost_{11}^{-\theta} & Cost_{12}^{-\theta} & \cdots & Cost_{1n}^{-\theta} \\ Cost_{21}^{-\theta} & Cost_{22}^{-\theta} & \cdots & Cost_{2n}^{-\theta} \\ \vdots & \vdots & \ddots & \vdots \\ Cost_{n1}^{-\theta} & Cost_{n2}^{-\theta} & \cdots & Cost_{nn}^{-\theta} \end{vmatrix}$$

where n is the number of cities in the data and θ is the trade elasticity (Donaldson and Hornbeck, 2016).[15]

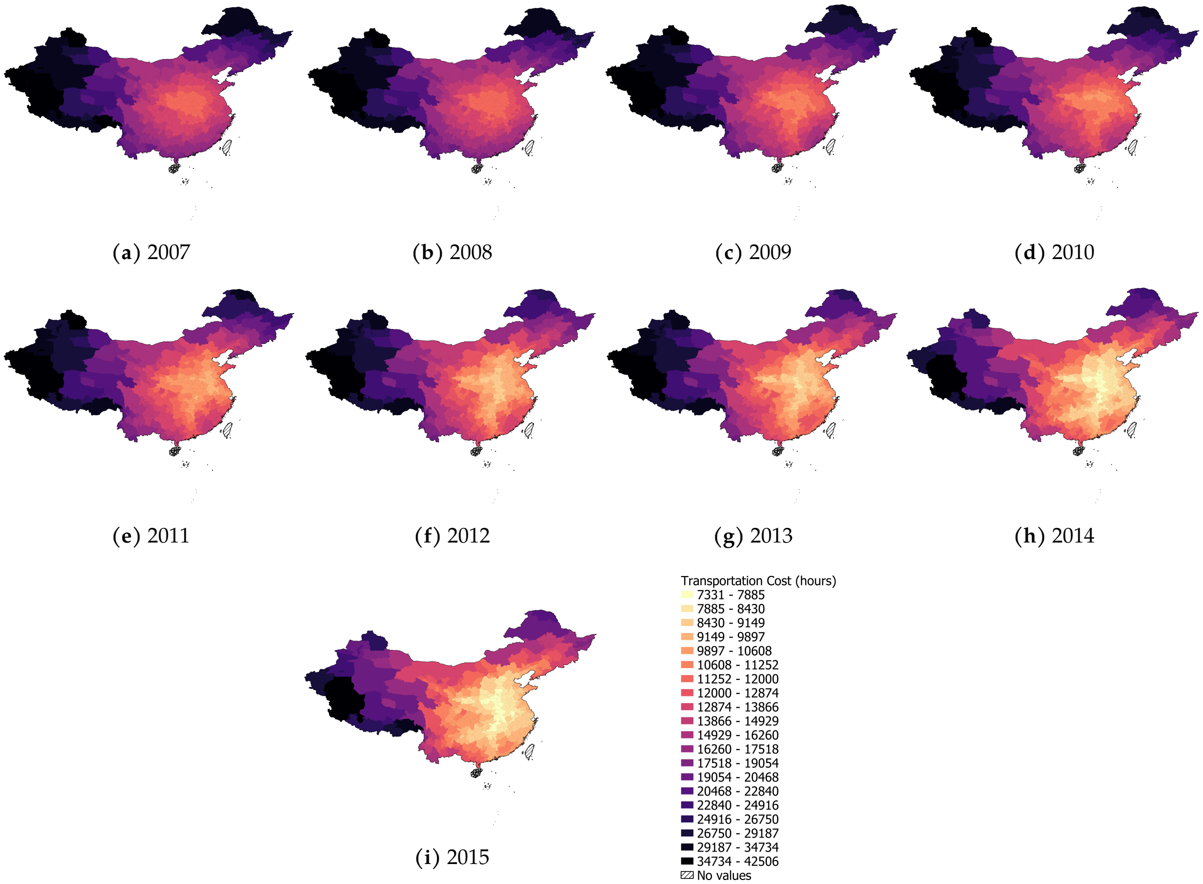
The purpose of the Figure 4 is to illustrate the connectivity level of Chinese cities. For instance, the light yellow cities of central China have a value between 7331 and 7885. This value reflects the hours needed for a given passenger from these cities to access all the other cities in China. As it can be observed, the HSR gradually improves the position of all the cities within the network. It facilitates the access even to the cities that are far away from central China - though the cities in the center South-East of the country are always better connected.

4.4 Other Variables

We use several control variables in our regressions. First, in order to control for economic activity at the city level, we rely on night light data (Henderson et al., 2011; Mellander

[15] For our empirical exercise we assume that θ is 1 but in the Appendix we also test the value of 8.22 used in Donaldson and Hornbeck (2016) and the value 12.86 used in Eaton and Kortum (2002), see Tables C.10 and C.11 in the Appendix C.

Figure 4: Transportation costs



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

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

(**i**) 2015

Notes: Authors' own computation.

et al., 2015), based on the harmonized global nighttime light dataset for the period 1992-2018 (Li et al., 2020). Second, we extract a large set of control variables from the HYDE database (Goldewijk et al., 2017).[16] It 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. We also include a dummy variable indicating if the city has national-level high-tech zones approved by the State Council, as in the official website of the Ministry of Science and Technology (MOST).[17]

In addition, we geo-localize flight passenger data at the airport level as an alternative transportation network (Wong, 2019), by using the Civil Aviation Administration of China website, to obtain data regarding the number of passengers. We then do reverse geo-coding from the names of the airports and aggregate the passenger data at the city level.

[16] We use version 3.2.

[17] http://www.most.gov.cn/zxgz/gxjscykfq/gxjsgxqml/, last accessed May 2021.

Table B.1 in the appendix separates the variables we use in each model and presents their summary statistics.

5 Results

5.1 The role of the HSR on innovation performance

We analyse the role of the HSR on the innovation performance of Chinese cities through a two-stage procedure, as commented in section 3. Panel A in table 1 contains the OLS results and panel B the IV estimations. Both OLS and IV coefficients for the distance to HSR are significant and negative, meaning that, for a given city, a reduction in the distance to the closest HSR station is associated with an increase in the number of patents per capita. To facilitate the interpretation of the coefficients, we standardized the variables in the regressions by subtracting the mean and dividing every value by its standard deviation. Thus, one standard deviation decrease in the distance to a HSR station results to approximately 0.24 standard deviations increase in patents per capita (panel B, column 1). Table B.2 in the appendix contains the IV first stage estimates. The positive effect of the instrument on the distance to the HSR station implies that having a courier station during the Ming dynasty increases the likelihood to be close to a HSR station in recent times.

In the second column of table 1, we remove first tier cities, according to the definition of Fang et al. (2015), and in the final column we keep only third tier cities.[18] The IV coefficients are slightly smaller than with the full sample of cities, but similar in magnitude and significance, showing that our effects are not driven solely by star cities, which are the most productive and innovative of the country. In all columns, we observe that IV coefficients are higher than OLS ones. In line with Dong et al. (2020), we explain it because the Chinese government intentionally planned some HSR stations in lagging areas to help them grow, which could explain the downward bias in the OLS estimates. As for the control variables, night light activity and passengers by air have a positive and significant effect on innovation activity, as expected.

5.2 Robustness analysis

In this section, we depart from the former OLS and IV regressions and run a number of additional estimations, in order to ensure that our results are robust to the choice of different

[18] The first tier cities are Beijing, Shanghai, Guangzhou, and Shenzhen.

Table 1: Main Results: HSR and Innovation

Dep. var. =	Patents per capita		
	(1)	(2)	(3)
Panel A	**OLS**	**OLS**	**OLS**
Log Distance to Stations	-0.103***	-0.106***	-0.112***
	[0.033]	[0.033]	[0.034]
Average Night Light	0.935***	0.931***	0.872***
	[0.195]	[0.201]	[0.213]
Flight Passengers	0.341***	0.453***	0.320**
	[0.088]	[0.109]	[0.130]
R-squared	0.83	0.78	0.76
Panel B	**IV**	**IV**	**IV**
Log Distance to Stations	-0.249***	-0.219**	-0.227***
	[0.086]	[0.087]	[0.087]
Average Night Light	0.860***	0.878***	0.817***
	[0.185]	[0.190]	[0.202]
Flight Passengers	0.320***	0.422***	0.289**
	[0.093]	[0.111]	[0.129]
First-Stage F-stat	19.22	19.09	19.91
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 population. Distance to HSR stations is computed as the straight line distance in meters from the most populous point of a city to the closest HSR station. Clustered standard errors at the city level are reported in brackets.

variables and methods of estimation.

We begin with a falsification exercise (Autor et al., 2013; Dix-Carneiro et al., 2018). We start by collecting data on population and patents from 2000 until 2007 from the same sources as in the benchmark analysis. Then, we run the IV regressions in which our key independent variable is the distance to the closest HSR station, as in Table 1. Differently, the dependent variable in this regression is now the number of patents per capita from 2000 to 2007, one year before the actual arrival of the HSR network. Thus, we regress past changes in patents per capita on future changes in distance to the HSR network, which we expect not to be significant. Table 2 presents the results of this exercise. Column 1 contains the IV estimates of the benchmark analysis of Table 1 after we restrict our sample from 2009 to 2016 in order to have the same number of years as in our falsification test. In column 2, we report no significant effect of future HSR network on past patents per capita. Based on these results we verify that our main findings are not capturing a long-run common causal factor behind both the expansion of rail network and the rise of patent activity. According to the estimates of Table 2 this relationship was absent in the years immediately prior to our sample period.

Next, we explore a conventional difference-in-differences regression model with stag-

Table 2: HSR and Innovation - Falsification Exercise

Dep. var. =	Patents per capita	Past Patents per capita
	(1)	(2)
Log Distance to Stations	-0.362***	
	[0.116]	
Log Future Distance to Stations		0.028
		[0.102]
First-Stage F-stat	13.52	13.52
Sample Size	2792	2792
City FE	Yes	Yes
Year FE	Yes	Yes
Sample	Full	Full

Notes: The dependent variables are (i) the number of patents divided by population from 2009 to 2016 (column 1) and (ii) the past number of patents divided by population from 2000 to 2007 (column 2). Both columns include all control variables as in the benchmark analysis. Clustered standard errors at the city level are reported in brackets.

gered treatment to control for pre-trends related to innovation activities before the arrival of HSR lines, which could have been attributed to the HSR in our previous analysis. By doing this, we allow for a long period of time (2000-2007) in which no city had access to a HSR station. As long as there are no systematic changes over time except for treatment, differences can be interpreted as causal. Thus, our panel starts in 2000, and run until 2016, and we compare our key outcome variables within a city before and after treatment. Our focal variable (*Access*) is now a binary indicator, instead of the distance to a HSR station, which switches to 1 if a city gains access to a HSR station, while including all our baseline controls and fixed effects:

$$Y_{it} = \alpha + \beta Access_{it-1} + \gamma_i + \delta_t + \zeta X_{it} + \epsilon_{it} \tag{8}$$

As observed in Table 3, we find that accessibility to a HSR station has a significant effect on the number of patents per capita. This positive effect means that there is an increase in innovation performance after a city gains access to a HSR station, in line with our benchmark findings.

From now on, the remaining robustness analyses are relegated to the online appendix C, although we discuss some of the main findings here.

First, according to Lin (2017), the HSR network uses speeds over 200 km/h. As a robustness test we repeat the same regressions but including the stations with connections to lines of 200 km/h and above (and not higher than 250 km/h as done in the baseline estimation). As presented in table C.1, our results are not affected.

Second, we provide evidence that the (short) time lag chosen between explanatory and dependent variables does not drive the results (table C.2). While the size of the coefficients

Table 3: HSR and Innovation - Difference in Differences Model

Dep. var. =	Patents per capita		
	(1)	(2)	(3)
Access to a HSR Station	0.106***	0.101***	0.082***
	[0.028]	[0.028]	[0.028]
Average Night Light	1.089***	1.209***	1.233***
	[0.221]	[0.207]	[0.219]
Flight Passengers	0.566***	0.482***	0.324***
	[0.153]	[0.124]	[0.120]
R-squared	0.72	0.65	0.62
Sample Size	5933	5865	5644
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 population, 2000-2016. Clustered standard errors at the city level are reported in brackets.

somewhat weakens the more you separate dependent and independent variables, it stays positive and highly significant.

Next, following concerns about CNIPA's patents quality (Eberhardt et al., 2017), we recompute our dependent variable either weighting it by the forward citations received or the claims contained in the application (tables C.3 and C.4, respectively), or considering only the patents that are at the 50% top of the distribution by technology and year, according to the forward citations received within a five-year window (table C.5). While coefficients slightly change, results and conclusions remain unaltered in all three tables.

Fourth, we also run robustness analyses on our IV strategy. In particular, we repeat the OLS-IV estimation (table 1) but adding two additional controls related to the Ming dynasty, multiplied with time dummies. The first one is the share of people that took the entry exams in each Chinese city (Wendy et al., 2012). [19] The second is the average population growth for the Chinese cities during Ming dynasty (Goldewijk et al., 2017). With these controls we aim to block off the potential direct effect of the couriers stations during the Ming dynasty on current innovation, and therefore the instrument meets the exclusion restriction. Table C.6 presents our results, which again, are similar to those in the benchmark analysis.

Finally, in order to mitigate potential concerns about the placement of the courier stops along their routes, we use as an instrument the distance to the routes of the couriers instead of the distance to the courier stations. Tables C.7 and C.8 presents the first and second stage results, respectively. Again, the results are similar to the benchmark analysis.

[19] The dataset is about the civil-service examination system in Imperial China, administered for the purpose of selecting candidates for the state bureaucracy.

5.3 HSR and knowledge diffusion

Yet, more importantly, we aim to explore the extent to which the expansion of the HSR is connected to knowledge diffusion, which ultimately affects innovation (Arrow, 1962; Romer, 1986; Weitzman, 1998). Following similar approaches, in China and elsewhere, we first assess HSR's effects on diffusion by means of city-to-city gravity models of patent citations and patent collaborations (Dong et al., 2020; Hanley et al., 2021). Differently from previous studies, we do not only compute the travel time from city to city due to the HSR, but build a bilateral variable of the least-cost path, taking into account the expansion of the HSR as well as other transportation means (see section 4.3). The way in which we address this analysis and the subsequent results are presented in the online appendix D. Overall, we find that there is a negative relationship between transportation costs among city pairs and citations and co-patenting flows, as expected. The relationship, though, is not excessively strong and concentrates in pairs of cities located at short distances (less than 250 km). Possibly, the HSR mostly facilitates commuting and short visits, which in turn allow researchers and technologists to meet face-to-face more frequently, and exchange and recombine ideas prone to more innovation.

Yet, as detailed in sections 2 and 3, using patent citations and patent collaborations present a number of drawbacks, some specific to the Chinese case, some other intrinsic to the underlying data. Thus, we rely on the branching literature to study whether cities' technological diversification into new domains is driven by their connection to experienced cities in that domain, thanks to the HSR, which we interpret as evidence of city to city knowledge diffusion (for similar approaches, see Bahar et al. 2020; Iasio and Miguelez 2022; Miguelez and Morrison 2022). Table 4 summarizes the results of the diversification model, which includes city-year and technology-year fixed effects, and controls for the number of patents that a city has in a given technology. The latter tries to mitigate concerns related to the fact that we can observe a certain technological specialization in a city because it has a very low number of patents which are concentrated in a given technology. We further control for the relatedness density and RTA of a city following the existing literature (Balland et al., 2019; Iasio and Miguelez, 2022). Column 1 reports strong evidence that, for a given city i and a given technology c, a higher access to technological knowledge produced by other cities already specialized in technology c affects positively the probability of the city i to specialize in technology c, too. The impact becomes higher when we control for the number of patents and the RTA value of a city in column (3).

In columns 4 to 6 of table 4 we rerun our main regression, but keeping time-invariant the number of patents in technology c as well as the dummy D_RTA_{jc} in the computation of the EEK index (both computed for the time window 1998-2007). The purpose of this exercise is to keep these two components of the EEK index constant, and thus all changes of our focal explanatory variable should come from solely the reduction in transportation costs due to the HSR expansion, allowing us to better gauge its impact on city-to-city diffusion. Overall, we observe that the probability of a city to specialize in a new technological field is associated to the specialization patterns of the cities to which the city connects to through the HSR. Also, note that the relatedness density index has a positive effect, which is in line with previous literature.

Table 4: Technological specialization model

Dep. var. =	Entry					
	(1)	(2)	(3)	(4)	(5)	(6)
Log (EEK+1)	0.052***	0.038***	0.062***			
	[0.005]	[0.005]	[0.004]			
Log (Time Invariant EEK+1)				0.025***	0.015***	0.035***
				[0.004]	[0.004]	[0.003]
Relatedness Density	0.014***	0.013***	0.013***	0.014***	0.013***	0.013***
	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]
Patents		-0.019***			-0.018***	
		[0.004]			[0.004]	
RTA		2.054***			2.169***	
		[0.080]			[0.080]	
Log (Patents+1)			-0.075***			-0.073***
			[0.002]			[0.002]
Log (RTA+1)			0.227***			0.229***
			[0.006]			[0.005]
R-squared	0.10	0.10	0.11	0.10	0.10	0.11
Sample Size	1631595	1631595	1631595	1631595	1631595	1631595
City*Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Year*IPC FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The dependent variable is binary and switches to 1 if a city has a comparative advantage in a specific technological field. Clustered standard errors at the city level are reported in brackets.

Despite the controls and fixed effects included in the regressions, we run a number of additional analyses in order to assess the robustness of our results.[20] First, table C.9 in appendix C repeats our analysis, but with slightly superior time lags between the dependent and our focal explanatory variable. Results are in line with expectations. Second, tables C.10 and C.11 repeat the analysis with two different values of θ (8.22 and 12.86, respectively), following the related literature (Donaldson and Hornbeck, 2016; Eaton and Kortum, 2002). θ aims to capture the space-related frictions on bilateral economic activity, with high values indicating that cities further away from a focal city i will have less impact on city i outcomes. As expected, when larger values of θ are considered, the impact of HSR expansion on knowledge diffusion becomes stronger, reflecting the fact that

[20] Unfortunately, an equivalent instrument to the one of the couriers' routes during the Ming dynasty is not possible to compute, among other things because the EEK is composed of three different variables. Other alternatives, such as historical/geographical links between cities, do not generally meet the necessary exclusion restrictions.

HSR increases the number of face-to-face interactions and exchanges at relatively short distances. Finally, we exclude from our analysis first- and second-tier cities in table C.12 in the appendix C. Results remain untouched when first-tier cities are excluded, while the coefficient decreases considerably (though remains strongly significant) when removing also second-tier cities.

6 Conclusion

This paper explores the role of the HSR expansion on innovation activity in China. Our analysis contributes to the growing literature on the effect of transportation networks on innovation activity (Agrawal et al., 2017; Andersson et al., 2021; Cui et al., 2020; Inoue and Nakajima, 2017; Perlman, 2015; Tamura, 2017), and particularly those studies assessing the roll-out of the HSR. In order to create exogenous variation for our analysis, we rely on historical couriers' stations (during the Ming dynasty) as a novel instrument. All in all, we report evidence that connectivity to HSR enhances the patenting activity of a city. Our results remain highly significant even when we remove the largest urban centers and innovative star cities, indicating that the effect on innovation is also driven by second and third-tier cities.

Secondly, we provide evidence on what is, we belief, one of the main mechanisms behind the HSR-innovation relationship, that is, knowledge diffusion and idea recombination. Thus, we build on the branching literature and look at the technological diversification of Chinese cities (Bahar et al., 2020; Balland et al., 2019; Petralia et al., 2017; Rigby, 2015) by presenting evidence that the probability of a city to diversify into a new technological field is related to the specialization patterns of the cities to which the city connects to through the HSR. Similarly, we also contribute to a recent streams of studies looking at the role of external inputs on the technological diversification paths of cities, which has been less studied in the literature (Elekes et al., 2019; Miguelez and Morrison, 2022; Neffke et al., 2017; Whittle et al., 2020).

Implications from our analysis are manifold and go beyond our empirical results. First, the idea that present times are characterized by a R&D productivity slowdown because the generation and flow of new ideas is becoming more costly is hotly debated (Bloom et al., 2020). Thus, new and better ideas are getting harder to find, not only in the US, but also in other economies, including China (Boeing and Hünermund, 2020). Among the causes, researchers points to an increasing need of specialists capable to go deeper in producing new knowledge, which in turn creates the need of larger and more varied teams of researchers in order to achieve the same level of idea complexity (Jones, 2009). Transportation infrastructure facilitates interactions and team formation, specially among people located far apart, thus potentially tackling the R&D productivity slowdown.

Second, our paper also contributes to the debate around income inequality. Some studies show that transport and educational infrastructures reduce personal income inequality, for instance in US states (Hooper et al., 2018). Further, historical analyses of the rail expansion concluded that the rail network allowed to move goods more easily, which fostered concentration in the core regions of countries (Krugman, 1991). However, the roll-out of the HSR, which reduces commuting and interaction costs, offers advantages to lessen spatial concentration (Büchel and Kyburz, 2020), and might smooth the consistent inequalities among mega cities and places that are left behind (**?**).

However, it must be acknowledged that transportation networks may also lead to personal inequality and concentration of economic resources too, as they may contribute to channel physical and human capital and resources to the more dynamic cities (Glaeser, 2011; Iammarino et al., 2019; Puga, 2002). The results in this paper need to be interpreted as the impact of a HSR network, which is a very specific infrastructure, with very specific people having access to it. According to Dobruszkes et al. (2022), the profile of HSR passengers is anything but neutral. HSR is more likely to be used by specific social groups, such as men in their thirties to fifties, with high income, high occupational positions and high educational levels. This is in line with findings that HSR network facilitates accessibility to the general rail network and has an uneven impact on the growth of Chinese cities (Jiao et al., 2020; Jin et al., 2020). This prevents us from generalizing our results to any kind of transportation infrastructure.

Acknowledgements

The paper was previously circulated as "Rails and innovation: Evidence from China". We would like to express our gratitude to Ron Boschma, Riccardo Crescenzi, Giovanni Dosi, Francesco Lissoni, Alex Coad, Antonio Andreoni, Carolina Castaldi, Gaétan de Rassenfosse, Bart Los, Jouke van Dijk, Ruben Gaetani, Martijn Smit, Yuan Tian, Nicola Cortinovis, Michele Pezzoni, Zhiling Wang, Valentina Di Iasio, 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. Finally, we would like to thank Wu Zhao and Yuan Quan for being research assistants during the process of data collecting and cleaning. Rosina Moreno and Ernest Miguelez acknowledge financial support provided by the Ministerio de Ciencia e Innovación for the project entitled 'Innovation, Environment and Social Justice (JUST GREEN-INN), PID2020-119096RB-I00. Georgios Tsiachtsiras acknowledges financial support from Agaur (FI-DGR). Any remaining errors are ours.

References

Aghion, P., Akcigit, U., Bergeaud, A., France, B. D., Blundell, R. and Hemous, D. (2019), 'Innovation and top income inequality', *Review of Economic Studies* **86**, 1–45.

Agrawal, A., Galasso, A. and Oettl, A. (2017), 'Roads and innovation', *Review of Economics and Statistics* **99**, 417–434.

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

Andersson, D., Berger, T. and Prawitz, E. (2021), 'Making a market: Infrastructure, integration, and the rise of innovation', *Review of Economics and Statistics* pp. 1–44.

Arora, A., Belenzon, S. and Lee, H. (2018), 'Reversed citations and the localization of knowledge spillovers', *Journal of Economic Geography* **18**, 495–521.

Arrow, K. J. (1962), 'The economic implications of learning by doing', *Review of Economic Studies* **29**, 155–173.

Aschauer, D. A. (1989), 'Is public expenditure productive?', *Journal of Monetary Economics* **23**, 177–200.

Autor, D. H., Dorn, D. and Hanson, G. H. (2013), 'The china syndrome: Local labor market effects of import competition in the united states', *American Economic Review* **103**, 2121–68.

Bahar, D., Choudhury, P. and Rapoport, H. (2020), 'Migrant inventors and the technological advantage of nations', *Research Policy* **49**, 103947.

Bahar, D., Hausmann, R. and Hidalgo, C. A. (2014), 'Neighbors and the evolution of the comparative advantage of nations: Evidence of international knowledge diffusion?', *Journal of International Economics* **92**, 111–123.

Balassa, B. (1965), 'Trade liberalisation and "revealed" comparative advantage', *Manchester School* **33**, 99–123.

Balland, P. A., Boschma, R., Crespo, J. and Rigby, D. L. (2019), 'Smart specialization policy in the european union: relatedness, knowledge complexity and regional diversification', *Regional Studies* **53**, 1252–1268.

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

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**, 435–448.

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

Beckers, V., Poelmans, L., Rompaey, A. V. and Dendoncker, N. (2020), 'The impact of urbanization on agricultural dynamics: a case study in belgium', *Journal of Land Use Science* **15**, 626–643.

Bloom, N., Jones, C. I., van Reenen, J. and Webb, M. (2020), 'Are ideas getting harder to find?', *American Economic Review* **110**, 1104–1144.

Boeing, P. and Hünermund, P. (2020), 'A global decline in research productivity? evidence from china and germany', *Economics Letters* **197**, 109646.

Boschma, R. (2017), 'Relatedness as driver of regional diversification: a research agenda', *Regional Studies* **51**, 351–364.

Bottasso, A., Conti, M., Robbiano, S. and Santagata, M. (2022), 'Roads to innovation: Evidence from italy', *Journal of Regional Science* **62**, 981–1005.

Brook, T. (1998), *The Confusions of Pleasure: Commerce and Culture in Ming China*, University of California Press.

Büchel, K. and Kyburz, S. (2020), 'Fast track to growth? railway access, population growth and local displacement in 19th century switzerland', *Journal of Economic Geography* **20**, 155–195.

Cao, Z., Derudder, B., Dai, L. and Peng, Z. (2021), ''buzz-and-pipeline' dynamics in chinese science: the impact of interurban collaboration linkages on cities' innovation capacity', *Regional Studies* pp. 1–17.

Carlino, G. and Kerr, W. (2014), 'Agglomeration and innovation', *NBER Working Paper 20367* .

Cui, J., Li, T. and Wang, Z. (2020), 'High-speed railway and collaborative innovation: Evidence from university patents in china', *SSRN Electronic Journal* .

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

Dix-Carneiro, R., Soares, R. R. and Ulyssea, G. (2018), 'Economic shocks and crime: Evidence from the brazilian trade liberalization', *American Economic Journal: Applied Economics* **10**, 158–95.

Dobruszkes, F., Chen, C.-L., Moyano, A., Pagliara, F. and Endemann, P. (2022), 'Is high-speed rail socially exclusive? an evidence-based worldwide analysis', *Travel Behaviour and Society* **26**, 96–107.

Donaldson, D. and Hornbeck, R. (2016), 'Railroads and american economic growth: A "market access" approach', *Quarterly Journal of Economics* **131**, 799–858.

Dong, X. (2018), 'High-speed railway and urban sectoral employment in china', *Transportation Research Part A: Policy and Practice* **116**, 603–621.

Dong, X., Zheng, S. and Kahn, M. E. (2020), 'The role of transportation speed in facilitating high skilled teamwork across cities', *Journal of Urban Economics* **115**, 103212.

Duan, L., Niu, D., Sun, W. and Zheng, S. (2021), 'Transportation infrastructure and capital mobility: evidence from china's high-speed railways', *Annals of Regional Science* **67**, 617–648.

Duranton, G. and Turner, M. A. (2012), 'Urban growth and transportation', *Review of Economic Studies* **79**, 1407–1440.

Eaton, J. and Kortum, S. (2002), 'Technology, geography, and trade', *Econometrica* **70**, 1741–1779.

Eberhardt, M., Helmers, C. and Yu, Z. (2017), 'What can explain the chinese patent explosion?', *Oxford Economic Papers* **69**, 239–262.

Elekes, Z., Boschma, R. and Lengyel, B. (2019), 'Foreign-owned firms as agents of structural change in regions', *Regional Studies* **53**, 1603–1613.

Essletzbichler, J. (2015), 'Relatedness, industrial branching and technological cohesion in us metropolitan areas', *Regional Studies* **49**, 752–766.

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

Fang, H., Gu, Q., Xiong, W. and Zhou, L.-A. (2015), 'Demystifying the chinese housing boom', *NBER Working Paper 21112* .

Feldman, M. P. and Kogler, D. F. (2010), 'Chapter 8 - stylized facts in the geography of innovation', **1**, 381–410.

Fogel, R. W. (1962), 'A quantitative approach to the study of railroads in american economic growth: A report of some preliminary findings*', *Journal of Economic History* **22**, 163–197.

Gao, J., Jun, B., Pentland, A. Zhou, T. and Hidalgo, C. A. (2021), 'Spillovers across industries and regions in china's regional economic diversification', *Regional Studies* pp. 1–16.

Gao, Y., Song, S., Sun, J. and Zang, L. (2018), 'Does high-speed rail really promote economic growth? evidence from chinaas yangtze river delta region', *SSRN Electronic Journal* .

Gao, Y. and Zheng, J. (2020), '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* **127**, 104838.

Garcia-Mila, T. and McGuire, T. J. (1992), 'The contribution of publicly provided inputs to states' economies', *Regional Science and Urban Economics* **22**, 229–241.

Glaeser, E. L. (2011), 'Triumph of the city : how our greatest invention makes us richer, smarter, greener, healthier, and happier', *New York: Penguin Press* p. 338.

Golden, M. and Min, B. (2013), 'Distributive politics around the world', *Annual Review of Political Science* **16**, 73–99.

Goldewijk, K. K., Beusen, A., Doelman, J. and Stehfest, E. (2017), 'Anthropogenic land use estimates for the holocene - hyde 3.2', *Earth System Science Data* **9**, 927–953.

Guo, Q. and He, C. (2017), 'Production space and regional industrial evolution in china', *GeoJournal* **82**, 379–396.

Hanley, D., Li, J. and Wu, M. (2021), 'High-speed railways and collaborative innovation', *Regional Science and Urban Economics* p. 103717.

Hausmann, R., Hwang, J. and Rodrik, D. (2007), 'What you export matters', *Journal of Economic Growth* **12**, 1–25.

He, C., Zhu, S. and Yang, X. (2017), 'What matters for regional industrial dynamics in a transitional economy?', *Area Development and Policy* **2**, 71–90.

Henderson, V., Storeygard, A. and Weil, D. N. (2011), A bright idea for measuring economic growth, Vol. 101, pp. 194–199.

Heuermann, D. F. and Schmieder, J. F. (2019), 'The effect of infrastructure on worker mobility: evidence from high-speed rail expansion in germany', *Journal of Economic Geography* **19**, 335–372.

Hidalgo, C. A., Balland, P. A., Boschma, R., Delgado, M., Feldman, M., Frenken, K., Glaeser, E. L., He, C., Kogler, D. F., Morrison, A., Neffke, F., Rigby, D., Stern, S., Zheng, S. and Zhu, S. (2018), 'The principle of relatedness'.

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

Hooper, E., Peters, S. and Pintus, P. A. (2018), 'To what extent can long-term investments in infrastructure reduce inequality?', *Journal of Infrastructure, Policy and Development* **2**, 193–225.

Hsiao, C., Ching, H. S. and Wan, S. K. (2012), '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* **27**, 705–740.

Huang, Y. and Wang, Y. (2020), 'How does high-speed railway affect green innovation efficiency? a perspective of innovation factor mobility', *Journal of Cleaner Production* **265**, 121623.

Iammarino, S., Rodriguez-Pose, A. and Storper, M. (2019), 'Regional inequality in europe: evidence, theory and policy implications', *Journal of Economic Geography* **19**, 273–298.

Iasio, V. D. and Miguelez, E. (2022), 'The ties that bind and transform: knowledge remittances, relatedness and the direction of technical change', *Journal of Economic Geography* **22**, 423–448.

Inoue, H. and Nakajima, K. (2017), 'The impact of the opening of high-speed rail on innovation', *RIETI Discussion Paper Series 17-E-034* .

Jaffe, A. B. and de Rassenfosse, G. (2017), 'Patent citation data in social science research: Overview and best practices', *Journal of the Association for Information Science and Technology* **68**, 1360–1374.

Jaffe, A. B., Trajtenberg, M. and Henderson, R. (1993), 'Geographic localization of knowledge spillovers as evidenced by patent citations', *Quarterly Journal of Economics* **108**, 577–598.

Jiao, J., Wang, J., Zhang, F., Jin, F. and Liu, W. (2020), 'Roles of accessibility, connectivity and spatial interdependence in realizing the economic impact of high-speed rail: Evidence from china', *Transport Policy* **91**, 1–15.

Jin, M., Lin, K. C., Shi, W., Lee, P. T. and Li, K. X. (2020), 'Impacts of high-speed railways on economic growth and disparity in china', *Transportation Research Part A: Policy and Practice* **138**, 158–171.

Jones, B. F. (2009), 'The burden of knowledge and the "death of the renaissance man": Is innovation getting harder?', *Review of Economic Studies* **76**, 283–317.

Ke, X., Chen, H., Hong, Y. and Hsiao, C. (2017), 'Do china's high-speed-rail projects promote local economy?—new evidence from a panel data approach', *China Economic Review* **44**, 203–226.

Komikado, H., Morikawa, S., Bhatt, A. and Kato, H. (2021), 'High-speed rail, inter-regional accessibility, and regional innovation: Evidence from japan', *Technological Forecasting and Social Change* **167**, 120697.

Krugman, P. R. (1991), *Geography and trade*, Leuven University Press.

Kulkarni, R., Haynes, K. E., Stough, R. R. and Riggle, J. D. (2012), '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* .

Lawrence, M., Bullock, R. and Liu, Z. (2019), 'China's high-speed rail development. international development in focus', *World Bank* .

Li, X., Zhou, Y., Zhao, M. and Zhao, X. (2020), 'A harmonized global nighttime light dataset 1992–2018', *Scientific Data* **7**.

Li, Y. (2016), 'China high speed railways and stations', *Harvard Dataverse, V1* .

Lin, Y. (2017), 'Travel costs and urban specialization patterns: Evidence from china's high speed railway system', *Journal of Urban Economics* **98**, 98–123.

Lin, Y., Qin, Y. and Zie, Z. (2015), 'International technology transfer and domestic innovation: Evidence from the high-speed rail sector in china', *CEP Discussion Papers* .

Lucas, R. E. (1993), 'Making a miracle', *Econometrica* **61**, 251.

Ma, Y., Su, H., Jin, Q., Feng, W., Liu, J. and Huang, W. (2016), *The General History of Chinese Tourism Culture*, SCPG PUBLISHING CORPORATION.

Melander, E. (2020), 'Transportation technology, individual mobility and social mobilisation', *CAGE Online Working Paper Series 471* .

Mellander, C., Lobo, J., Stolarick, K. and Matheson, Z. (2015), 'Night-time light data: A good proxy measure for economic activity?', *PLOS ONE* **10**, e0139779.

Miguelez, E. and Morrison, A. (2022), 'Migrant inventors as agents of technological change', *Journal of Technology Transfer* pp. 1–24.

Mimeur, C., Queyroi, F., Banos, A. and Thévenin, T. (2018), 'Revisiting the structuring effect of transportation infrastructure: An empirical approach with the french railway network from 1860 to 1910', *Historical Methods* **51**, 65–81.

Moretti, E. (2021), 'The effect of high-tech clusters on the productivity of top inventors', *American Economic Review* **111**, 3328–75.

Munnell, A. H. (1992), 'Policy watch: Infrastructure investment and economic growth', *Journal of Economic Perspectives* **6**, 189–198.

Neffke, F., Hartog, M., Boschma, R. and Henning, M. (2017), 'Agents of structural change: The role of firms and entrepreneurs in regional diversification', *Economic Geography* **94**, 23–48.

Owen-Smith, J. and Powell, W. W. (2004), 'Knowledge networks as channels and conduits: The effects of spillovers in the boston biotechnology community', *Organization Science* **15**, 5–21.

Perlman, E. R. (2015), 'Dense enough to be brilliant: Patents, urbanization, and transportation in nineteenth century america', *CEH Discussion Papers* .

Petralia, S., Balland, P. A. and Morrison, A. (2017), 'Climbing the ladder of technological development', *Research Policy* **46**, 956–969.

Puga, D. (2002), 'European regional policies in light of recent location theories', *Journal of Economic Geography* **2**, 373–406.

Qin, Y. (2017), "no county left behind?' the distributional impact of high-speed rail upgrades in china', *Journal of Economic Geography* **17**, 489–520.

Redding, S. and Turner, M. A. (2015), 'Transportation costs and the spatial organization of economic activity'.

Rigby, D. L. (2015), 'Technological relatedness and knowledge space: Entry and exit of us cities from patent classes', *Regional Studies* **49**, 1922–1937.

Romer, P. M. (1986), 'Increasing returns and long-run growth', *Journal of Political Economy* **94**, 1002–1037.

Soete, L. (1987), 'The impact of technological innovation on international trade patterns: The evidence reconsidered', *Research Policy* **16**, 101–130.

Sun, D., Zeng, S., Ma, H. and Shi, J. J. (2021), 'How do high-speed railways spur innovationx003f;', *IEEE Transactions on Engineering Management* .

Tamura, R. (2017), 'The effect of high-speed railways on knowledge transfer: Evidence from japanese patent citations', *Public Policy Review* **13**, 325–342.

The Economist (2018), 'China's tech industry is catching up with silicon valley'.

The Economist (2019), 'America still leads in technology, but china is catching up fast'.

Tsiachtsiras, G. (2020), 'Transportation networks and the rise of the knowledge economy in 19th century france article', *University of Barcelona mimeo* .

Tubiana, M., Miguelez, E. and Moreno, R. (2022), 'In knowledge we trust: Learning-by-interacting and the productivity of inventors', *Research Policy* **51**, 104388.

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

Veugelers, R. (2017), 'The challenge of china's rise as a science and technology powerhouse', *Policy Contributions* .

Wang, X. and Feng, Y. (2021), 'The effects of national high-tech industrial development zones on economic development and environmental pollution in china during 2003–2018', *Environmental Science and Pollution Research* **28**, 1097–1107.

Weitzman, M. L. (1998), 'Recombinant growth', *The Quarterly Journal of Economics* **113**, 331–360.

Wendy, W., Bol, P. K., Lewis, B. G., Bertrand, M., Berman, M. L., Blossom, J. C. and Guan, W. W. (2012), 'Worldmap-a geospatial framework for collaborative research', *Annals of GIS* **18**, 121–134.

Whittle, A., Lengyel, B. and Kogler, D. F. (2020), 'Understanding regional branching knowledge diversification via inventor collaboration networks', *Papers in Evolutionary Economic Geography (PEEG)* **13**, 51–68.

WIPO (2019), 'World intellectual property indicators 2019'.

WIPO (2021), 'World intellectual property indicators 2021'.

Wong, J. C. Y. (2019), 'Blue-sky thinking: Connectivity impacts on regional economies and innovation in the united states *'.

Wuchty, S., Jones, B. F. and Uzzi, B. (2007), 'The increasing dominance of teams in production of knowledge', *Science* **316**, 1036–1039.

Yang, X., Zhang, H., Lin, S., Zhang, J. and Zeng, J. (2021), 'Does high-speed railway promote regional innovation growth or innovation convergence?', *Technology in Society* **64**, 101472.

Yao, L. and Li, J. (2022), 'Intercity innovation collaboration and the role of high-speed rail connections: evidence from chinese co-patent data', *Regional Studies* pp. 1–13.

Yin, D., Motohashi, K. and Dang, J. (2020), 'Large-scale name disambiguation of chinese patent inventors (1985–2016)', *Scientometrics* **122**, 765–790.

Yu, L. (2021), 'Study on treatment effects and spatial spillover effects of beijing–shanghai hsr on the cities along the line', *Annals of Regional Science* **67**, 671–695.

Zhang, Y., Wu, B., Tan, L. and Liu, J. (2021), 'Quantitative research on the efficiency of ancient information transmission system: A case study of wenzhou in the ming dynasty', *PLoS ONE* **16**.

Zheng, S. and Kahn, M. E. (2013), '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* **110**, E1248–E1253.

Zhu, S., He, C. and Luo, Q. (2019), 'Good neighbors, bad neighbors: local knowledge spillovers, regional institutions and firm performance in china', *Small Business Economics* **52**, 617–632.

Zhuang, L. and Ye, C. (2020), 'Changing imbalance: Spatial production of national high-tech industrial development zones in china (1988-2018)', *Land Use Policy* **94**, 104512.

Zou, W., Chen, L. and Xiong, J. (2019), 'High-speed railway, market access and economic growth', *International Review of Economics and Finance* .

A Technical Details on Variable Construction

A.1 Transportation Cost

A HSR line with an average speed of 300 kilometers per hour is assigned a value of 120, for a HSR line with 250 kilometers per hour the value of 144, for a HSR line with 200 kilometers per hour the value of 180, for a general rail line with a speed of 130 kilometers per hour the value of 277, for a highway line with an average speed of 110 kilometers per hour the value 327, for a national road line with an average speed of 90 kilometers per hour the value of 400, for the other smaller roads the value of 600 which is based on the average travel speed of 60 kilometers per hour and for the grids that are not crossed by any line, we follow the existing literature (Mimeur et al., 2018) and we assign to them the value of 7200 based on the 5 kilometers per hour walking speed. The cost values reflect the necessary seconds for crossing a grid. The travel speeds for the road network are assigned based on the speed limits in Asia.[1] Regarding the general rail network, since we do not have information about the speed of each line, we assign a speed of 130 kilometers per hour, which is the average speed of all the different train lines that exist in China.[2] Then, based on the type of line that crosses each grid, we allocate one of the values 120, 144, 180, 277, 327. 400 or 600. If no line is intersected with a grid then the grid takes the value of 7200. Over the period 2007-2015, the Chinese government dedicated much of its effort to the construction of the HSR network. Since the other networks did not evolve much over time, most of their effect should be captured by the fixed effects in the regression analysis.

A.2 Relatedness Density Index

To compute the relatedness density index, we first calculate the co-occurrence of any two different technologies (following the IPC classification codes) in the same patent document. In other words, we count how many times any two technologies appear together in a same patent document (co-occurrence) by using all the patents available in the dataset.[3] We control for the fact that this co-occurrence can be random by normalizing our measure using the association measure by van Eck and Waltman (2009), resulting in a co-occurrence

[1] https://www.wikiwand.com/en/Speed_limits_in_China, last accessed May 2022.

[2] https://www.wikiwand.com/en/Rail_transport_in_China, last accessed May 2022.

[3] We take the four-digit, second level of disaggregation of IPC, in order to get a total number of 600 technological classes for the period from 1985 to 2016.

measure of technology c and any other technology d:

$$\phi_{cd} = \frac{o_{cd}}{e_{cd}} \tag{A.1}$$

where ϕ_{cd} stands for relatedness between technology c and technology d, o_{cd} for the normalized co-occurrence of those two technologies and e_{cd} for the expected co-occurrence. The expected co-occurrences are computed by taking into account the number of times technology c and d occur in our database during a given period (s_c and s_d respectively) relative to the total number of patents in the same period, p. This yields the final formula:

$$\phi_{cd} = (\frac{o_{cd}}{s_c s_d}) * p. \tag{A.2}$$

Second, to analyse how relatedness influences technological diversification at the city level, we have to construct a city-level variable that indicates how close one technology is to the existing portfolio of technologies of a given city. Thus, to transfer the measure of co-occurrence obtained above to the city level, we combine the RTA_{ict} measure given in equation (5) with the co-occurrence matrix obtained above to know the extent to which each technology c is related to a city's existing technologies (or city's knowledge portfolio) through the following density index (Balland et al., 2015):

$$RelDen_{ict} = (\frac{\sum_{c \epsilon r} \phi_{ict} RTA_{ict}}{\sum_{c} \phi_{ict}}) x100 \tag{A.3}$$

This density index computes how related is the technological base of a city i at time t with respect to technology c. The intuition behind is that two technologies are highly related to each other when they are frequently observed at the same time in a city. And this degree of relatedness can influence the specialization of the city in technology c. This is why we want to control for this fact, so that the effect attributed to the reduction in transportation cost thanks to the HSR is not contaminated by the influence of this relatedness index.

B Additional tables

Table B.1: Summary statistics

Variables	Obs	Mean	Std. Dev.	Min	Max
Model of innovation performance					
Patents per 10,000 People	3141	2.134	4.837	0	54.12
Past Patents per 10,000 People	2792	.266	.914	0	23.583
Distance to HSR stations with more than 250 average speed	3141	670647.8	894601	864.921	4857225
Distance to Ming stations	3141	2271454	4921132	322.239	2.48e+07
Average night light area	3141	6.911	8.119	.001	58.336
Air transportation (passengers)	3141	1617290	7272339	0	1.00e+08
High Technological Zones	3141	.245	.43	0	1
Cropland Per Capita	3141	.001	.001	0	.009
Grazing Per Capita	3141	.022	.168	0	2.945
Difference-in-Differences Model					
Patents per 10,000 People	5933	1.255	3.694	0	54.12
Access to a HSR station	5933	.131	.338	0	1
Average night light area	5933	6.063	7.618	.001	58.336
Air transportation (passengers)	5933	1325430	6309836	0	1.08e+08
High Technological Zones	5933	.213	.41	0	1
Cropland Per Capita	5933	.001	.001	0	.009
Grazing Per Capita	5933	.022	.167	0	2.945
Model of Technological Specialization					
Entry	1631595	.088	.284	0	1
Exposure to External Knowledge	1631595	2.992	14.254	0	1277.732
Time Invariant Exposure to External Knowledge	1631595	1.135	5.003	.001	490.407
Relatedness density	1631595	5.425	16.745	0	100
Number of Patents per IPC Class	1631595	1.506	22.379	0	4739
RTA	1631595	.051	.175	0	1

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

Dep. var. =	Log Distance to Stations		
	(1)	(2)	(3)
Log Distance to Courier Station X 2009	0.205***	0.199***	0.197***
	[0.032]	[0.031]	[0.029]
Log Distance to Courier Station X 2010	0.474***	0.463***	0.450***
	[0.049]	[0.049]	[0.046]
Log Distance to Courier Station X 2011	0.582***	0.577***	0.559***
	[0.053]	[0.053]	[0.052]
Log Distance to Courier Station X 2012	0.522***	0.520***	0.515***
	[0.054]	[0.054]	[0.052]
Log Distance to Courier Station X 2013	0.553***	0.550***	0.555***
	[0.057]	[0.057]	[0.055]
Log Distance to Courier Station X 2014	0.601***	0.600***	0.612***
	[0.056]	[0.056]	[0.054]
Log Distance to Courier Station X 2015	0.482***	0.480***	0.491***
	[0.071]	[0.072]	[0.073]
Log Distance to Courier Station X 2016	0.428***	0.425***	0.439***
	[0.072]	[0.073]	[0.075]
Average Night Light	-0.485***	-0.448***	-0.460***
	[0.101]	[0.102]	[0.104]
Flight Passengers	-0.119***	-0.206**	-0.166
	[0.045]	[0.088]	[0.154]
Sample Size	3141	3105	2988
City FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes

Notes: Clustered standard errors at the city level are reported in brackets.

C Robustness analysis

Table C.1: HSR stations with speed higher than 200km per hour and Innovation

Dep. var. =	Patents per capita		
	(1)	(2)	(3)
Panel A	**OLS**	**OLS**	**OLS**
Log Distance to Stations	-0.109***	-0.111***	-0.126***
	[0.035]	[0.035]	[0.035]
Average Night Light	0.943***	0.941***	0.879***
	[0.196]	[0.202]	[0.214]
Flight Passengers	0.339***	0.443***	0.310**
	[0.089]	[0.112]	[0.128]
R-squared	0.84	0.78	0.76
Panel B	**IV**	**IV**	**IV**
Log Distance to Stations	-0.250***	-0.219**	-0.223**
	[0.087]	[0.088]	[0.087]
Average Night Light	0.885***	0.903***	0.843***
	[0.187]	[0.193]	[0.205]
Flight Passengers	0.316***	0.406***	0.279**
	[0.093]	[0.120]	[0.131]
First-Stage F-stat	22.77	22.22	22.37
Sample Size	3141	3105	2988
City FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Sample	Full	Second and Third	Third

Notes: Clustered standard errors at the city level are reported in brackets.

Table C.2: HSR and Innovation - Time Lags

Dep. var. =	Patents per capita			
	(1)	(2)	(3)	(4)
Panel A	**OLS**	**OLS**	**OLS**	**OLS**
Lag 1 Log Distance to Stations	-0.103***			
	[0.033]			
Lag 2 Log Distance to Stations		-0.072***		
		[0.021]		
Lag 3 Log Distance to Stations			-0.093***	
			[0.023]	
Lag 4 Log Distance to Stations				-0.090***
				[0.020]
R-squared	0.83	0.86	0.89	0.91
Panel B	**IV**	**IV**	**IV**	**IV**
Lag 1 Log Distance to Stations	-0.249***			
	[0.086]			
Lag 2 Log Distance to Stations		-0.141***		
		[0.053]		
Lag 3 Log Distance to Stations			-0.130**	
			[0.053]	
Lag 4 Log Distance to Stations				-0.144***
				[0.048]
First-Stage F-stat	19.22	19.95	23.69	26.28
Sample Size	3141	2792	2443	2094
City FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Sample	Full	Full	Full	Full

Notes: The dependent variable is the number of patents divided by population. Distance to HSR stations contains the straight line distance in meters from the most populous point of a city to the closest HSR station with a speed more than 250km per hour. Average night light data controls for the economic activity of a city. Flight passengers accounts for the effect of alternative transportation networks. We control for the cropland and grazing density of the city and we include a dummy variable switching to 1 if the city belongs to a high technological zone. The variable distance to a HSR station has been transformed using the log 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 brackets.

Table C.3: HSR and Innovation - Inventions weighted by their Citations in a 5 year window

Dep. var. =	Citations 5 year window per capita		
	(1)	(2)	(3)
Panel A	**OLS**	**OLS**	**OLS**
Log Distance to Stations	-0.077***	-0.078***	-0.070***
	[0.021]	[0.019]	[0.020]
Average Night Light	0.645***	0.632***	0.587***
	[0.130]	[0.131]	[0.140]
Flight Passengers	0.310***	0.289***	0.191**
	[0.106]	[0.072]	[0.084]
R-squared	0.90	0.84	0.81
Panel B	**IV**	**IV**	**IV**
Log Distance to Stations	-0.230***	-0.204***	-0.175***
	[0.069]	[0.066]	[0.065]
Average Night Light	0.567***	0.573***	0.536***
	[0.132]	[0.131]	[0.141]
Flight Passengers	0.288***	0.254***	0.163*
	[0.109]	[0.071]	[0.085]
First-Stage F-stat	19.22	19.09	19.91
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 weighted by their number of citations within a 5 year window divided by population. Clustered standard errors at the city level are reported in brackets.

Table C.4: HSR and Innovation - Inventions weighted by their Claims

Dep. var. =	Claims per capita		
	(1)	(2)	(3)
Panel A	**OLS**	**OLS**	**OLS**
Log Distance to Stations	-0.074***	-0.074***	-0.072***
	[0.022]	[0.020]	[0.021]
Average Night Light	0.781***	0.729***	0.674***
	[0.168]	[0.167]	[0.179]
Flight Passengers	0.271***	0.291***	0.204**
	[0.092]	[0.069]	[0.091]
R-squared	0.87	0.80	0.78
Log Distance to Stations	-0.180***	-0.154**	-0.144**
	[0.064]	[0.065]	[0.069]
Average Night Light	0.727***	0.692***	0.639***
	[0.171]	[0.168]	[0.182]
Flight Passengers	0.256***	0.269***	0.185*
	[0.093]	[0.071]	[0.095]
Panel B	**IV**	**IV**	**IV**
First-Stage F-stat	19.22	19.09	19.91
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 weighed by the number of claims divided by population. Clustered standard errors at the city level are reported in brackets.

Table C.5: HSR and Innovation - Top Cited Patents in a 5 year window

Dep. var. =	Top Cited Patents per capita		
	(1)	(2)	(3)
Panel A	**OLS**	**OLS**	**OLS**
Log Distance to Stations	-0.083***	-0.086***	-0.094***
	[0.028]	[0.029]	[0.029]
Average Night Light	0.862***	0.854***	0.813***
	[0.186]	[0.192]	[0.203]
Flight Passengers	0.320***	0.436***	0.315**
	[0.073]	[0.117]	[0.135]
R-squared	0.84	0.77	0.74
Panel B	**IV**	**IV**	**IV**
Log Distance to Stations	-0.186**	-0.153**	-0.169**
	[0.073]	[0.074]	[0.075]
Average Night Light	0.810***	0.823***	0.777***
	[0.180]	[0.185]	[0.197]
Flight Passengers	0.305***	0.418***	0.295**
	[0.077]	[0.119]	[0.135]
First-Stage F-stat	19.22	19.09	19.91
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 count of patents most cited (top-50 percent) in a five years window, divided by population. Clustered standard errors at the city level are reported in brackets.

Table C.6: HSR and Innovation - Additional Controls during Ming Dynasty

Dep. var. =	Patents per capita		
	(1)	(2)	(3)
Log Distance to Stations	-0.101***	-0.104***	-0.111***
	[0.033]	[0.033]	[0.034]
Average Night Light	0.920***	0.917***	0.865***
	[0.196]	[0.202]	[0.215]
Flight Passengers	0.340***	0.452***	0.321**
	[0.089]	[0.109]	[0.131]
R-squared	0.84	0.78	0.76
Log Distance to Stations	-0.258***	-0.224**	-0.234**
	[0.091]	[0.092]	[0.091]
Average Night Light	0.848***	0.868***	0.812***
	[0.187]	[0.192]	[0.204]
Flight Passengers	0.318***	0.420***	0.288**
	[0.094]	[0.112]	[0.130]
First-Stage F-stat	17.43	17.24	18.04
Sample Size	3141	3105	2988
City FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Ming Controls	Yes	Yes	Yes
Sample	Full	Second and Third	Third

Notes: Clustered standard errors at the city level are reported in brackets.

Table C.7: HSR and Innovation - Routes of Couriers as an Instrument

Dep. var. =	Log Distance to Stations		
	(1)	(2)	(3)
Log Distance to Courier Routes X 2009	0.143***	0.139***	0.139***
	[0.028]	[0.028]	[0.027]
Log Distance to Courier Routes X 2010	0.319***	0.311***	0.304***
	[0.037]	[0.037]	[0.036]
Log Distance to Courier Routes X 2011	0.392***	0.388***	0.380***
	[0.041]	[0.041]	[0.040]
Log Distance to Courier Routes X 2012	0.347***	0.346***	0.347***
	[0.041]	[0.041]	[0.040]
Log Distance to Courier Routes X 2013	0.367***	0.365***	0.374***
	[0.044]	[0.044]	[0.043]
Log Distance to Courier Routes X 2014	0.397***	0.396***	0.411***
	[0.043]	[0.043]	[0.043]
Log Distance to Courier Routes X 2015	0.327***	0.326***	0.339***
	[0.050]	[0.051]	[0.052]
Log Distance to Courier Routes X 2016	0.296***	0.295***	0.310***
	[0.051]	[0.052]	[0.053]
Average Night Light	-0.483***	-0.447***	-0.457***
	[0.102]	[0.103]	[0.105]
Flight Passengers	-0.114***	-0.195**	-0.141
	[0.044]	[0.090]	[0.154]
Sample Size	3141	3105	2988
City FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes

Notes: Clustered standard errors at the city level are reported in brackets.

Table C.8: HSR and Innovation - Routes of Couriers as an Instrument: Second Stage

Dep. var. =	Patents per capita		
	(1)	(2)	(3)
Log Distance to Stations	-0.234**	-0.204**	-0.223**
	[0.094]	[0.096]	[0.097]
Average Night Light	0.868***	0.885***	0.818***
	[0.185]	[0.190]	[0.201]
Flight Passengers	0.322***	0.426***	0.290**
	[0.094]	[0.113]	[0.130]
First-Stage F-stat	14.32	14.07	14.27
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 population. Distance to HSR stations is computed as the straight line distance in meters from the most populous point of a city to the closest HSR station. Clustered standard errors at the city level are reported in brackets.

Table C.9: Technological specialization model - Lags

Dep. var. =	Entry			
	(1)	(2)	(3)	(4)
Lag 1 Log (EEK+1)	0.062***			
	[0.004]			
Lag 2 Log (EEK+1)		0.057***		
		[0.004]		
Lag 3 Log (EEK+1)			0.052***	
			[0.004]	
Lag 4 Log (EEK+1)				0.048***
				[0.004]
Relatedness Density	0.013***	0.013***	0.013***	0.013***
	[0.001]	[0.001]	[0.001]	[0.001]
Log (Patents+1)	-0.075***	-0.074***	-0.073***	-0.071***
	[0.002]	[0.002]	[0.002]	[0.002]
Log (RTA+1)	0.227***	0.229***	0.229***	0.229***
	[0.006]	[0.006]	[0.006]	[0.006]
R-squared	0.11	0.10	0.10	0.10
Sample Size	1631595	1442519	1255475	1070650
City*Year FE	Yes	Yes	Yes	Yes
Year*IPC FE	Yes	Yes	Yes	Yes

Notes: The dependent variable is binary and switches to 1 if a city has a comparative advantage in a specific technological field. Clustered standard errors at the city level are reported in brackets.

Table C.10: Technological specialization model - Elasticity of Donaldson and Hornbeck

Dep. var. =	Entry					
	(1)	(2)	(3)	(4)	(5)	(6)
Log (EEK+1) DH	0.050***	0.050***	0.045***			
	[0.005]	[0.005]	[0.005]			
Log (Time Invariant EEK+1) DH				0.069***	0.056***	0.074***
				[0.005]	[0.005]	[0.004]
Relatedness Density	0.002***	0.002***	0.002***	0.014***	0.013***	0.013***
	[0.000]	[0.000]	[0.000]	[0.001]	[0.001]	[0.001]
Patents		-0.001			-0.019***	
		[0.003]			[0.004]	
RTA		0.374***			2.052***	
		[0.062]			[0.076]	
Log (Patents+1)			-0.051***			-0.072***
			[0.003]			[0.002]
Log (RTA+1)			0.087***			0.225***
			[0.005]			[0.005]
R-squared	0.31	0.31	0.31	0.10	0.10	0.11
Sample Size	1629435	1629435	1629435	1631595	1631595	1631595
City*Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Year*IPC FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The dependent variable is binary and switches to 1 if a city has a comparative advantage in a specific technological field. We apply the trade elasticity of 8.22 to our EEK measure following the value proposed by Donaldson and Hornbeck (2016). Clustered standard errors at the city level are reported in brackets.

Table C.11: Technological specialization model - Elasticity of Eaton and Kortum

Dep. var. =	Entry					
	(1)	(2)	(3)	(4)	(5)	(6)
Log (EEK+1) EK	0.050***	0.049***	0.044***			
	[0.005]	[0.005]	[0.005]			
Log (Time Invariant EEK+1) EK				0.078***	0.065***	0.081***
				[0.005]	[0.005]	[0.004]
Relatedness Density	0.002***	0.002***	0.002***	0.014***	0.013***	0.013***
	[0.000]	[0.000]	[0.000]	[0.001]	[0.001]	[0.001]
Patents		-0.001			-0.019***	
		[0.003]			[0.004]	
RTA		0.379***			2.053***	
		[0.062]			[0.075]	
Log (Patents+1)			-0.051***			-0.072***
			[0.003]			[0.002]
Log (RTA+1)			0.086***			0.225***
			[0.004]			[0.005]
R-squared	0.31	0.31	0.31	0.10	0.10	0.11
Sample Size	1629435	1629435	1629435	1631595	1631595	1631595
City*Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Year*IPC FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The dependent variable is binary and switches to 1 if a city has a comparative advantage in a specific technological field. We apply the trade elasticity of 12.86 to our EEK measure following the value proposed by Eaton and Kortum (2002). Clustered standard errors at the city level are reported in brackets.

Table C.12: Technological specialization model - Second and Third Tier Cities

Dep. var. =	Entry			
	(1)	(2)	(3)	(4)
Log (EEK+1)	0.064***		0.068***	
	[0.004]		[0.004]	
Log (Time Invariant EEK+1)		0.036***		0.038***
		[0.003]		[0.003]
Relatedness Density	0.012***	0.012***	0.012***	0.012***
	[0.001]	[0.001]	[0.001]	[0.001]
Log (Patents+1)	-0.077***	-0.074***	-0.079***	-0.076***
	[0.002]	[0.002]	[0.003]	[0.002]
Log (RTA+1)	0.224***	0.226***	0.220***	0.222***
	[0.006]	[0.005]	[0.006]	[0.006]
R-squared	0.11	0.11	0.11	0.11
Sample Size	1616228	1616228	1569166	1569166
City*Year FE	Yes	Yes	Yes	Yes
Year*IPC FE	Yes	Yes	Yes	Yes
Sample	Second and Third	Second and Third	Third	Third

Notes: The dependent variable is binary and switches to 1 if a city has a comparative advantage in a specific technological field. Clustered standard errors at the city level are reported in brackets.

D Cross-city citations and collaborations

In the second part of our empirical analysis, we investigated one potential mechanism on the relationship between the HSR roll-out and the city-level innovation, that is, the possibility of knowledge spreading, which ultimately affects the recombination of ideas and the production of new knowledge. Our approach in the main part of the paper builds on the branching literature, and estimates the likelihood of specializing in a new technology as a function of the increased connections to other cities. Yet, most of the literature looking at cross-city knowledge diffusion has addressed the issue using city-to-city patent citations, as well as counts of co-patents.

First, co-patenting has been widely used as a measure of teamwork and team collaboration, which has been shown to lead to knowledge diffusion between team members. Breschi and Lissoni (2009) empirically demonstrate that networks of co-inventors strongly affect the diffusion of ideas across firms and space. They find solid evidence that the most important reason why geography matters in constraining the diffusion of knowledge is that it limits the interactions (collaborations) among inventors. Akcigit et al. (2018) find that inventors' collaborations facilitate individuals' knowledge accumulation, thanks to frequent interactions with their peers.

Second, since the work of Jaffe et al. (1993), patent citations have been used as a measure of knowledge diffusion, as they have been considered a valid paper trail of knowledge spillovers. Although whether patent citations can reflect real knowledge flows has been under intense debate (Alcacer and Gittelman, 2006; Corsino et al., 2019; Jaffe and de Rassenfosse, 2017), it is acknowledge as the best way available to capture knowledge spillovers and has been widely used in many studies (Buzard et al., 2020; Diemer and Regan, 2022; Kwon et al., 2020; Murata et al., 2014).

We proceed by estimating gravity equations of the following form:

$$Know_{ijt} = \alpha + \beta Cost_{ijt-1} + \xi_{ij} + \rho_{it} + \zeta_{jt} + \epsilon_{ijt} \tag{D.1}$$

where $Know_{ijt}$ is the aggregated number of co-applications or citations across two cities in year t, and where i and j refer to origin and destination city, respectively. $Cost_{ijt-1}$ is the transportation cost based on the HSR lines across two cities, computed for the period $t-1$. We include origin - destination fixed effects, ξ_{ij}, origin - year fixed effects, ρ_{it}, and destination - year fixed effects, ζ_{jt}. We estimate by OLS and cluster the standard errors at the level of origin and destination cities.[4]

A specific limitation of the use of CNIPA data to measure collaborations is that it does not contain information of inventors' addresses, but only the address of the first applicant (Yin et al., 2020), thus making impossible to compute measures of co-inventorship across cities. We assume that inventors listed in a patent application with more than one applicant belong to different firms and therefore constitute teams that work across firms. Therefore, we use cross-city co-applications to proxy for collaborations. We fill in the address of non-first applicants with their corresponding information provided in other patents. For instance, one applicant may not be the first applicant in a given patent application a but it can be the first one in another patent application b. Then, we observe the address of this applicant from the patent application b. We then rely on the assumption that firms do not change their address frequently as to fill in the missing values in the case of non-first applicants with their address in the closest observable year in the dataset. After filling in the addresses, we geo-code the data and we identify 490,314 co-application cases, of which 260,030 (about 53.03%) are cross-city collaborations.

First, Table D.2 presents the regression results of the role of transportation cost on co-applications based on equation D.1. We obtain that the cities which are close in terms of distance (better connected thanks to the HSR) are the ones that are most benefited by the HSR network in terms of patent co-applications. In Table D.3, we repeat the same exercise with USPTO patent collaborations, showing evidence again that the effect of transportation on collaborations is driven by the cities that are within the given threshold of 250 km.

Table D.1: Summary statistics

Variables	Obs	Mean	Std. Dev.	Min	Max
Gravity Model Co-applications					
Patent Co-applications	546534	.217	5.428	0	1248
USPTO Patent Collaborations	546534	.012	.769	0	216
Total transportation cost in minutes	546534	2520.197	1708.785	9.233	11152.21
Transportation cost for cities within 250 km	546534	6.692	50.988	0	2457.956
Transportation cost for cities from 250 to 499 km	546534	35.351	151.223	0	3161.543
Transportation cost for cities from 500 to 749 km	546534	80.715	280.793	0	4074.127
Transportation cost for cities from 750 to 999 km	546534	127.895	400.52	0	4723.083
Transportation cost for cities from 1000 to 1249 km	546534	165.383	511.341	0	5462.423
Gravity Model Citations					
Total number of citations	1086804	17.692	227.754	0	44257
Citations within a 5 year window	1086804	15.109	200.745	0	39235
USPTO patent citations	1086804	.004	.216	0	75
Total transportation cost in minutes	1086804	2506.884	1701.036	9.233	11152.21
Transportation cost for cities within 250 km	1086804	6.71	51.047	0	2457.956
Transportation cost for cities from 250 to 499 km	1086804	35.424	151.254	0	3161.543
Transportation cost for cities from 500 to 749 km	1086804	81.071	281.283	0	4074.127
Transportation cost for cities from 750 to 999 km	1086804	128.599	401.483	0	4723.083
Transportation cost for cities from 1000 to 1249 km	1086804	166.155	512.167	0	5462.423

[4] We rely on the OLS estimation method of Correia (2015) with high-dimensional fixed effects.

Table D.2: Transportation cost and patent co-applications - OLS

Dep. var. =	Co-applications	
	(1)	(2)
Standardized values of lltotalcost3	-0.053**	
	[0.023]	
Less than 250 km		-0.736***
		[0.188]
250-499 km		-0.239
		[0.150]
500-749 km		-0.126*
		[0.072]
750-999 km		-0.050
		[0.079]
1000-1249 km		-0.228
		[0.227]
More than 1250 km		0.105
		[0.137]
R-squared	0.68	0.68
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 D.1 estimated with OLS and standard errors clustered at the origin and destination city.

Second, we explore the effect of transportation costs on the number of citations between patents in different cities. Table D.4 summarizes these results. In line with the ones based on co-applications, we observe that the effect is basically due to the cities that are within a short distance. To be more specific, even though in Panel A we do not find a general significant effect of the HSR network on the citations between cities, we report evidence in Panel B that the diffusion process based on citation data is driven by the cities that are within less than 749 km of distance. Our results are confirmed also when we use subsequent years in columns 2 and 3. We repeat the same exercise in Table D.5, using USPTO patent citation data and we find a similar effect. Finally, in D.6 we restrict our analysis only to the citations within a 5-year window and again the same pattern is observed.

Table D.3: Transportation cost and USPTO patent collaborations - OLS

Dep. var. =	USPTO Patent Collaborations	
	(1)	(2)
Standardized values of lltotalcost3	-0.010	
	[0.011]	
Less than 250 km		-0.263***
		[0.066]
250-499 km		-0.018
		[0.036]
500-749 km		0.072
		[0.052]
750-999 km		0.071
		[0.065]
1000-1249 km		0.081
		[0.074]
More than 1250 km		-0.106
		[0.122]
R-squared	0.66	0.66
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 D.1 estimated with OLS and standard errors clustered at the origin and destination city.

Table D.4: Transportation cost and patent citations - OLS

Dep. var. =	Citations	Citations t+1	Citations t+2
	(1)	(2)	(3)
Log Transportation Cost	-0.0398	-0.0333	-0.0277
	[0.0288]	[0.0282]	[0.0235]
R-squared	0.88	0.92	0.94
Less than 250 km	-0.7517***	-0.6140***	-0.3436***
	[0.2047]	[0.1633]	[0.0990]
250-499 km	-0.2663***	-0.2303***	-0.1699***
	[0.0850]	[0.0798]	[0.0642]
500-749 km	-0.0934*	-0.0909*	-0.0867**
	[0.0564]	[0.0541]	[0.0434]
750-999 km	-0.0041	-0.0104	-0.0374
	[0.0596]	[0.0584]	[0.0506]
1000-1249 km	-0.1926	-0.1648	-0.1282
	[0.1320]	[0.1214]	[0.0897]
More than 1250 km	0.0204	0.0238	-0.0179
	[0.1716]	[0.1692]	[0.1408]
R-squared	0.88	0.92	0.94
Sample Size	1086804	1086804	1086804
City Origin x Destination FE	Yes	Yes	Yes
Destination FE x Year FE	Yes	Yes	Yes
Origin FE x Year FE	Yes	Yes	Yes

Notes: Gravity model based on equation D.1 estimated with OLS and standard errors clustered at the origin and destination city.

Table D.5: Transportation cost and USPTO patent citations - OLS

Dep. var. =	USPTO Citations	
	(1)	(2)
Log Transportation Cost	-0.0157	
	[0.0171]	
Less than 250 km		-0.6698***
		[0.2361]
250-499 km		0.0781
		[0.0514]
500-749 km		0.0868*
		[0.0473]
750-999 km		0.0692
		[0.0640]
1000-1249 km		0.0130
		[0.0526]
More than 1250 km		-0.0544
		[0.1434]
R-squared	0.52	0.52
Sample Size	1086804	1086804
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 D.1 estimated with OLS and standard errors clustered at the origin and destination city.

Table D.6: Transportation cost and patent citations within a 5 year window - OLS

Dep. var. =	Cit5	Cit5 t+1	Cit5 t+2
	(1)	(2)	(3)
Log Transportation Cost	-0.0503	-0.0404	-0.0295
	[0.0353]	[0.0326]	[0.0250]
R-squared	0.85	0.90	0.94
Less than 250 km	-0.8751***	-0.6904***	-0.3574***
	[0.2385]	[0.1828]	[0.1040]
250-499 km	-0.3154***	-0.2662***	-0.1826***
	[0.1012]	[0.0915]	[0.0692]
500-749 km	-0.1160*	-0.1094*	-0.0960**
	[0.0666]	[0.0622]	[0.0480]
750-999 km	-0.0166	-0.0223	-0.0424
	[0.0708]	[0.0676]	[0.0558]
1000-1249 km	-0.2350	-0.1891	-0.1281
	[0.1590]	[0.1378]	[0.0921]
More than 1250 km	0.0011	0.0101	-0.0222
	[0.2087]	[0.1932]	[0.1487]
R-squared	0.85	0.90	0.94
Sample Size	1086804	1086804	1086804
City Origin x Destination FE	Yes	Yes	Yes
Destination FE x Year FE	Yes	Yes	Yes
Origin FE x Year FE	Yes	Yes	Yes

Notes: Gravity model based on equation D.1 estimated with OLS and standard errors clustered at the origin and destination city.

References

Akcigit, U., Soler, S. C., Miguelez, E., Stantcheva, S. and Sterzi, V. (2018), 'Dancing with the stars: Innovation through interactions', *NBER Working Paper 24466* .

Alcacer, J. and Gittelman, M. (2006), 'Patent citations as a measure of knowledge flows: The influence of examiner citations', *The Review of Economics and Statistics* **88**, 774–779.

Balland, P. A., Boschma, R. and Frenken, K. (2015), 'Proximity and innovation: From statics to dynamics', *Regional Studies* **49**, 907–920.

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

Buzard, K., Carlino, G. A., Hunt, R. M., Carr, J. K. and Smith, T. E. (2020), 'Localized knowledge spillovers: Evidence from the spatial clustering of rd labs and patent citations', *Regional Science and Urban Economics* **81**, 103490.

Correia, S. (2015), 'Singletons, cluster-robust standard errors and fixed effects: A bad mix *'.

Corsino, M., Mariani, M. and Torrisi, S. (2019), 'Firm strategic behavior and the measurement of knowledge flows with patent citations', *Strategic Management Journal* **40**, 1040–1069.

Diemer, A. and Regan, T. (2022), 'No inventor is an island: Social connectedness and the geography of knowledge flows in the us', *Research Policy* **51**, 104416.

Donaldson, D. and Hornbeck, R. (2016), 'Railroads and american economic growth: A "market access" approach', *Quarterly Journal of Economics* **131**, 799–858.

Eaton, J. and Kortum, S. (2002), 'Technology, geography, and trade', *Econometrica* **70**, 1741–1779.

Jaffe, A. B. and de Rassenfosse, G. (2017), 'Patent citation data in social science research: Overview and best practices', *Journal of the Association for Information Science and Technology* **68**, 1360–1374.

Jaffe, A. B., Trajtenberg, M. and Henderson, R. (1993), 'Geographic localization of knowledge spillovers as evidenced by patent citations', *Quarterly Journal of Economics* **108**, 577–598.

Kwon, H. S., Lee, J., Lee, S. and Oh, R. (2020), 'Knowledge spillovers and patent citations: trends in geographic localization, 1976–2015', *Economics of Innovation and New Technology* pp. 1–25.

Mimeur, C., Queyroi, F., Banos, A. and Thévenin, T. (2018), 'Revisiting the structuring effect of transportation infrastructure: An empirical approach with the french railway network from 1860 to 1910', *Historical Methods* **51**, 65–81.

Murata, Y., Nakajima, R., Okamoto, R. and Tamura, R. (2014), 'Localized knowledge spillovers and patent citations: A distance-based approach', *Review of Economics and Statistics* **96**, 967–985.

van Eck, N. J. and Waltman, L. (2009), 'How to normalize cooccurrence data? an analysis of some well-known similarity measures', *Journal of the American Society for Information Science and Technology* **60**, 1635–1651.

Yin, D., Motohashi, K. and Dang, J. (2020), 'Large-scale name disambiguation of chinese patent inventors (1985–2016)', *Scientometrics* **122**, 765–790.

UBIREA

Institut de Recerca en Economia Aplicada Regional i Pública
Research Institute of Applied Economics

WEBSITE: www.ub.edu/irea • CONTACT: irea@ub.edu

AQR

Grup de Recerca Anàlisi Quantitativa Regional
Regional Quantitative Analysis Research Group

WEBSITE: www.ub.edu/aqr/ • CONTACT: aqr@ub.edu

Universitat de Barcelona
Av. Diagonal, 690 • 08034 Barcelona